DOCUMENT RESUME

ED 230 618 TM 830 460

AUTHOR	Brandt, David A.; And Others
TITLE	Development of the National Assessment of Educational Progress.
INSTITUTION	American Institutes for Research in the Behavioral Sciences, Palo Alto, Calif.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
REPORT NO	AIR-25900-11/82-FR
PUB DATE	Nov 82
CONTRACT	NIE-400-82-0015
NOTE	87p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Educational Assessment; Educational Testing; Elementary Secondary Education; *Federal Programs; *Program Descriptions; *Program Development; *Program Improvement; Psychometrics
IDENTIFIERS	*National Assessment of Educational Progress

ABSTRACT

This report discusses five issue areas in which it is believed substantial improvements in the National Assessment of Educational Progress (NAEP) might be achieved. The unifying themes among these issues are to increase the visibility of NAEP, its relevance to policymakers, and its utility to state and local agencies. The first of five substantive chapters deals with the critical need for an overall framework for NAEP objectives. The second chapter deals with the design of test administration and focuses on the costs and benefits of a unified, integrated assessment given each year. The next chapter indicates that exercises are the wrong unit of analysis for NAEP and compares latent trait and latent class approaches to the development of meaning in assessment, pointing out the special applicability of the latent class analysis for achievement indices. The fourth chapter reviews existing studies of computer-based testing, seeks out predictions of future technological advances, and proposes a gradual series of studies aimed at the ultimate infusion of computer-administered tests throughout NAEP. In the final chapter, a concept and plan are described for an Educational Assessment Institute. Primary type of information provided by report: Program Description (Operating Policies); Procedures (Conceptual). (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

AIR-25900-11/82-FR

ED230618

AMERICAN INSTITUTES FOR RESEARCH
IN THE BEHAVIORAL SCIENCES
P.O. Box 1113, 1791 Arastradero Rd., Palo Alto, Ca. 94302 ● 415/493-3550

Development of the National Assessment of Educational Progress

David A. Brandt
John G. Claudy
Kevin J. Gilmartin
Steven M. Jung
Donald H. McLaughlin
Sandra R. Wilson

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

This work was done under Contract No. 400-82-0015 with the National Institute of Education, Department of Education. The content does not necessarily reflect the position or policy of either agency, however, and no official endorsement should be inferred.

November 1982

TM 830 460

vol. 2

An Equal Opportunity Employer

TABLE OF CONTENTS

	Page
Preface	iii
Objectives for a National Assessment of Educational Progress	1
The Role and History of NAEP Objectives	1
Current Assessment Objectives	2
Future NAEP Assessment Objectives	16
Annual Assessments of Learning Areas	31
General Specifications	32
Potential Benefits	34
Potential Costs	38
Conclusion	41
Measurement Founded on Modern Psychometric Theory	43
Background	43
Aptitude versus Achievement	44
Multiple Matrix Sampling	45
Reporting Results from Individual Exercises	47
Latent Trait Analysis	48
Latent Class Analysis	50
Conclusion	53
Computer-Administered Testing	55
Feasibility	55
Practicality	55
Recommended Feasibility Study	57
Conclusion	60
Establishment of an Educational Assessment Institute	61
Introduction	61
Potential Functions for an Educational Assessment Institute	61
Institute Organization	67
Institute Funding	67
Present Status of Institute Planning	68
References	69
Appendix A: PLAN Master Objectives (Foreword and Introduction)	71

LIST OF TABLES

		Page
Table 1.	Current Objectives of the National Assessment of Educational Progress	3
Table 2.	Summary of NAEP Assessment Schedule	14
Table 3.	Summary of Released Exercises and Estimates of Relative Numbers of Exercises Per Assessment by Assessment Area	15
Table 4.	Critical Factors Related to the Quality of Life of Adults	19
Table 5.	Educational Goals for Elementary and Secondary Education as Adopted by State Governments	21

Preface

The National Assessment of Educational Progress was established with several purposes, one of which was to pioneer new methods for filling information needs in education. In 1982, the National Institute of Education took such a pioneering step: the funding of five parallel projects aimed at producing plans for carrying out NAEP. This step was taken, at least in part, because of the dearth of proposals received in 1978 in response to the previous announcement of a NAEP grant competition.

And that step has been effective. The American Institutes for Research and (we assume) the Educational Testing Service and the National Opinion Research Center have prepared competing proposals for NAEP while exploring innovations that can improve NAEP no matter who the grantee may be. AIR (and we assume ETS and NORC) has reached the conclusion that the inadequacy with which NAEP has been performed by the incumbent demands a fresh approach.

In our proposal for the planning grant, AIR discussed six issue areas in which we believed substantial improvements in NAEP might be achieved. The unifying themes among these issues areas were increasing the visibility of NAEP, its relevance to policymakers, and its utility to state and local education agencies. This report expands five of those areas. We have not devoted a chapter to the sixth because, upon further consideration, the improvements appeared straightforward and not in the least problematic. Our efforts in each of the other five areas represent initial steps, to be expanded upon by the NAEP grantee. We feel each of the innovations we describe is essential for the future health and productivity of NAEP.

The first of the five substantive chapters of this report deals with objectives. Sandra Wilson, Director of AIR's Medical College Admissions Test effort, carefully examined the NAEP objectives and compared their characteristics with the objectives of MCAT, which she developed through a procedure that can be applied to NAEP. Dr. Wilson points out the critical need for an overall framework for NAEP objectives and outlines a plan for surmounting the difficulties engendered by bringing up the topic of

"refining objectives." The groundwork that John Flanagan laid, through Project PLAN and through his study of educational goals, provides a good starting point for this effort.

The second substantive chapter deals with the design of test administration and focuses on the costs and benefits of a unified, integrated assessment given each year. Don McLaughlin, Director of AIR's NAEP Planning Grant effort, describes one scenario for NAEP and points out that assessing each area each year (in each booklet, even) will improve the responsiveness, power, and research utility of NAEP at little or no cost. These conclusions are based on the experience of AIR in policy research and on the expertise of our Social Indicators Research Program staff.

In the next chapter, David Brandt, who has recently joined AIR from the University of Chicago Behavioral Science Department, agrees with Darrell Bock that exercises are the wrong unit of analysis for NAEP but goes further to compare latent trait and latent class approaches to the development of meaning in the assessment. While both have their place, David points out the special applicability of latent class analysis for achievement indices. In any case, it is essential, he argues, to modify the NAEP matrix sampling design to allow estimation of scores across booklets.

The fourth substantive chapter discusses the dream of computer-administered testing in down-to-earth terms. Dr. John Claudy, former Director of the Project TALENT Data Bank and currently a senior staff member in AIR's Measurement, Analysis, and Utilization Group, reviewed existing studies of computer-based testing, sought out predictions of future technological advances, and proposes a gradual series of studies aimed at the ultimate infusion of computer-administered tests throughout NAEP.

In the final chapter, Dr. Steven Jung, Director of AIR's Institute for Analysis of Educational Policy, describes AIR's concept and plans for an Educational Assessment Institute. We believe this to be a special aspect of AIR's planning for NAEP, and our partner in the development of

the concept, the Stanford University School of Education, has special qualifications for managing such an institute.

The chapter we did not develop, on modifications of the sampling design to make estimates for a variety of target groups and issues as they arise, was to have been written by Dr. Lauress Wise. He examined the issues involved and attested to the fact that there were really no important issues--of course it could be done. We did reach the conclusion that unless somebody else is busy producing a school district file from the 1980 Census, it might be appropriate for NAEP to tackle this effort for the school districts contained in sampled PSUs.

Others who worked on this effort include Paul Schwarz, President of AIR, William Clemans, Vice President and Director of AIR's Palo Alto office, Robert Krug, AIR's Director of Research, and senior research staff including Barbara Bessey, Bob Rossi, Darlene Russ-Eft, Terry Armstrong, Laurie Wise, and Patti Bourexis. John Flanagan provided guidance, especially for the discussion of objectives. Kevin Gilmartin edited the text and put together the planning grant report.

Donald H. McLaughlin

November 1, 1982

Objectives for a National Assessment of Educational Progress

The Role and History of NAEP Objectives

The basic purpose of the National Assessment of Educational Progress is to collect data and report over time on the performance of young Americans in reading, mathematics, and communications; to conduct assessments in other subject areas as the need arises; and to provide state and local educational agencies with technical assistance in interpreting assessment results and conducting their own assessments. In attempting to implement the first two purposes, NAEP selected ten major learning areas as its original focus, and within each of these areas it developed lists of specific objectives on which performance was to be assessed. As the assessment has developed, certain areas have been combined such that there are now eight assessment areas. The NAEP objectives form the heart of the assessment, and the set of objectives in each assessment area are the end result of extended deliberation by subject matter specialists, educators, and concerned lay persons over more than 17 years--as advisors, consultants to NAEP, contractors, or as NAEP staff. These objectives "form the framework for the learning area assessed" (NAEP, SY-OI-36), encompassing the knowledge, skills, understandings, and attitudes that are to be assessed in each area.

Because of their centrality, it is essential that these objectives be examined as they are now, after having been hammered out in a long, arduous, and expensive consensus process, and then we need to consider what issues have arisen and might arise with respect to the objectives in the future and how these might best be addressed.

Before considering the objectives themselves, however, it is important to realize several general features about them. The first is that the development of objectives has been carried out with a high degree of independence from one assessment area to the next. Hence, each set of objectives has been subject to the prevailing perspectives of experts in that content area as to the structure and the desirable level of detail and as to how objectives should be made appropriate for the different age

groups to be assessed. Within a given area, objectives (and exercises) have been required to meet at least one of the following criteria, namely, that they

- be considered important by scholars in the discipline in question,
- be acceptable to most educators as desirable teaching goals in most schools, and
- be considered desirable by thoughtful lay citizens.

A point worth noting is that these "criteria" are not so much criteria as they are a requirement that the ratification of certain constituent groups be obtained if an objective is to be included in any assessment. Another point to note is that the NAEP objectives have not gone without serious criticism. Greenbaum, Garet, and Solomon (1977), for example, have severely criticized NAEP's approach to setting objectives, noting that the discipline or subject-matter assessment approach has serious limitations. Furthermore, many concerns voiced during the original development process were never directly addressed but simply set aside as the highly political process surrounding the formulation of objectives went forward.

Current Assessment Objectives

NAEP's assessments of content areas have been performed separately, and similarly the assessment objectives have been developed and published separately. This practice makes it difficult to review the NAEP objectives as a whole and to interrelate objectives across content areas. Table 1 contains an outline of current ECS/NAEP assessment objectives based on the eight most recent NAEP objectives booklets:

- Reading and Literature Objectives: 1979-80 Assessment (1980)
- Writing Objectives: Second Assessment (1972)
- Mathematics Objectives: 1981-82 Assessment (1981)

Table 1

Current Objectives of the National Assessment of Educational Progress

READING AND LITERATURE

I. Values reading and literature

A. Values the benefits of reading for the individual

1. Recognizes that reading can be a source of enjoyment; demonstrates a commitment to reading for enjoyment

- Do students feel that some reading might be personally enjoyable?
- Do students identify reading, among other activities, as a source of enjoyment?
- Do students spend time reading for enjoyment? What do they read? How often?

2. Recognizes that written materials can contribute to personal growth; demonstrates a commitment to reading as one means of developing self-understanding

- Do students think they might learn about themselves and others through reading?
- Do students read for their own personal growth?

3. Recognizes that reading can be a means of acquiring knowledge and solving problems; demonstrates a commitment to reading as a means of acquiring knowledge and solving problems

- Do students think that reading might be a valuable source of information?
- Do students read to gain knowledge and solve problems?

B. Appreciates the cultural role of written discourse as a way of transmitting, sustaining, and changing the values of a society

- Do students recognize that written materials and society influence each other?
- Do students support the written expression of different viewpoints?

II. Comprehends written works

A. Comprehends words and lexical relationships

- Can students understand the meaning of words when used in the context of written material?
- Can students understand figurative and idiomatic meanings of words?
- Can students understand case relationships such as actor, action, and recipient?
- Can students understand anaphoric relationships between words and their referents?

B. Comprehends Propositional Relationships

- Can students understand propositional relationships—such as causality, temporality, and instrumentality—that are clearly stated in a paragraph?
- Can students understand propositional relationships—such as causality, temporality, and instrumentality—that are implied in a paragraph?

C. Comprehends Textual Relationships

- Can students infer the main idea or purpose of a text?
- Can students understand the character, mood, theme, or meaning in a text?
- Can students understand various explanations for states or events?

III. Responds to written works in interpretive and evaluative ways

A. Extends understanding of written works through interpretation

1. Demonstrates awareness of emotional impact of written works

- Do students experience emotion in responses to written works intended to be funny, sad, provocative, and so on?
- Can students relate their emotions to the purpose and meaning of a written work?

2. Applies personal experience to written works

- Do students recognize relationships between their own experience and something they read?
- Can students effectively apply personal experiences to what they read in order to deepen their understanding?

3. Applies knowledge of other works or other fields of study

- Can students relate what they read to other works?
- Can students relate knowledge of other fields of study, such as history, science, or philosophy, to what they read?

4. Analyzes written works

- Can students identify the formal structure of a work and see how that structure contributes to the meaning of the work?
- Can students identify literary devices and see how these devices contribute to the meaning of the work?

B. Evaluates Written Works

- What criteria do students use to evaluate poems and stories?
- Can students apply appropriate criteria to evaluate a broad range of written works?

IV. Applies study skills in reading

A. Obtains information from nonprose reading facilitators

- Do students use visual aids when reading?
- Can students correctly interpret information given on a chart, map, or graph?

B. Uses the various parts of a book

- Do students use different parts of a book to find information?
- Can students use the different parts of a book to find specific information?

C. Obtains information from materials commonly found in libraries or resource centers

- Do students use various reference materials?
- Can students find specific information from reference materials?

D. Uses various study techniques

- Do students use various techniques to aid their studying?
- Can students adjust their reading rates depending on their purpose for reading?

WRITING

I. Demonstrates ability to reveal personal feelings and ideas

A. Through free expression

B. Through the use of conventional modes of discourse

II. Demonstrates ability to write in response to a wide range of societal demands and obligations. Ability is defined to include correctness in usage, punctuation, spelling, and form or convention as appropriate to particular writing tasks, e.g., manuscripts, letters.

A. Social

1. Personal
2. Organizational
3. Community

B. Business/vocational

C. Scholastic

III. Indicates the importance attached to writing skills

A. Recognizes the necessity of writing for a variety of needs (as in I and II)

B. Writes to fulfill those needs

C. Gets satisfaction, even enjoyment, from having written something well

MATHEMATICS

COGNITIVE DOMAIN

I. Content

A. Numbers and numeration

1. Numeration (whole numbers, fractions, decimals, percent, integers, scientific notation)
2. Number concepts (whole numbers, fractions, decimals, percent, integers)
3. Operations (whole numbers, fractions, decimals, percents, integers)
4. Mental computation
5. Estimation
6. Properties
7. Relations

B. Variables and relationships

1. Facts, definitions, and symbols
2. Use of variables in equations and inequalities (solutions, equivalences, and translations)
3. Operations with variables
4. Use of variables to represent elements of a number system
5. Functions and formulas
6. Coordinate systems
7. Exponential and trigonometric functions
8. Logic

C. Shape, size, and position

1. Recognition of figures
2. Constructions and drawings
3. Visualization (static and dynamic)
4. Recognition of relationships (congruence, similarity, and symmetry)
5. Identification of properties from given visual information within, between, or among figures
6. Relationships involving classes of figures
7. Definitions, postulates, and theorems (recall, inference, and application)

D. Measurement

1. Unit (appropriate size and type of unit, unit equivalents, conversions within a system)

2. Instrument reading (English and metric rulers, scales, thermometers, clocks, etc.)

3. Linear measure (including nonstandard units)

4. Area, perimeter, and volume

5. Precision

6. Estimation of measurements

E. Probability and statistics

1. Organizing, displaying, and interpreting information (tallies, graphs, charts, and tables)

2. Measures of central tendency (mean, median, mode)

3. Measures of spread and position (range, percentile, standard deviations)

4. Sampling and polling

5. Probability (simple, compound, and independent events; odds)

6. Combinations and permutations

F. Technology

1. Hand calculator

2. Computer literacy

II. Process

A. Mathematical knowledge

- How well can students recall and recognize facts, definitions, and symbols?

B. Mathematical skill

- How well can students perform paper-and-pencil computation, including computations with whole numbers, integers, fractions, decimals, percents, and ratios and proportions?
- How well can students perform algebraic manipulations?
- How well can students perform geometric manipulations like constructions and spatial visualizations?
- How well can students make measurements?
- How well can students read graphs and tables?
- How well can students compute statistics, probabilities, or combinations?
- How well can students perform mental computations, including computation with whole numbers, fractions, decimals, and percents?
- How well can students estimate the answers to computations and measurements?
- How well can students perform computations involving whole numbers, decimals, fractions, and percents using calculators?
- How well can students read flow charts or basic computer programs?

C. Mathematical understanding

- How well can students translate a verbal statement into symbols or a figure, and vice versa?
- How well do students understand mathematical concepts and principles?
- How well can students select the appropriate uses of computers?
- How well can students select an appropriate computational method such as paper and pencil, mental, estimation, or calculator?

D. Mathematical application

- How well can students solve routine textbook problems?
- How well can students solve nonroutine problems?
- How well can students apply problem-solving strategies?
- How well can students interpret data and draw conclusions?
- How well can students use mathematics, including logic, in reasoning and making judgments?
- How well can students use a calculator to solve application problems?

AFFECTIVE DOMAIN

A. Attitudes

- How do students feel about the mathematics they encounter in school?
- How do students feel about the various activities in mathematics classes?
- How do students feel about their personal experience with mathematics?
- What are students' beliefs about the nature of mathematics as a discipline?
- What are students' beliefs about the value of mathematics to society?
- What are students' beliefs about computers?

CITIZENSHIP AND SOCIAL STUDIES

I. Demonstrates skills necessary to acquire information

- A. Uses the senses
- B. Uses sources such as card catalogues and indexes, case studies, computers, drawings, films, globes and other models, graphs, maps, newspapers, photos, pictures, radio, recordings, reference books, slides, tapes, television
- C. Uses techniques such as personal interviews, written essays, polls, and questionnaires

II. Demonstrates skills necessary to use information

- A. Organizes information
- B. Applies information
- C. Makes decisions and solves problems
- D. Critically evaluates information

III. Demonstrates an understanding of individual development and the skills necessary to communicate with others

- A. Examines individual beliefs, values, and behaviors
- B. Demonstrates individual development
- C. Communicates in graphic and oral forms
- D. Gives attention and reponds to the expression of others
- E. Interacts in groups in various capacities
- F. Has effective relations with people having different cultural perspectives

IV. Demonstrates an understanding of and interest in the ways human beings organize, adapt to, and change their environments

- A. Understands the forces that shape individual human beings
- B. Understands the interrelatedness of human societies
- C. Understands the organization of human societies
- D. Understands the relationships between individuals and groups
- E. Understands the relationships among groups
- F. Understands the relationships between people and the natural environment
- G. Has an awareness of global concerns
- H. Has a commitment to human rights worldwide

V. Demonstrates an understanding of and interest in the development of the United States

- A. Understands the principles and purposes of the United States
- B. Understands the organization and operation of the governments in the United States
- C. Understands political decision making in the United States
- D. Understands the electoral processes in the United States
- E. Understands the basis and organization of the legal system in the United States
- F. Knows rights of individuals in the United States
- G. Recognizes civil and criminal judicial systems in the United States
- H. Has a commitment to support justice and rights of all individuals
- I. Understands economics in the United States
- J. Understands major social changes that have occurred in American society
- K. Has a commitment to participating in community service and civic improvement

SCIENCE

COGNITIVE DOMAIN

I. Content

- A. Biology
  1. Germ theory and disease
  2. Systematics
  3. Cell theory
  4. Energy transformation
  5. Heredity
  6. Systems
  7. Evolution
  8. Ecology
  9. Behavior
  10. Growth and development
- B. Physical science
  1. Matter
  2. Combinations
  3. Mechanics
  4. Waves
  5. Electricity and magnetism
- C. Earth science
  1. Meterology
  2. Geology
  3. Oceanography
  4. Astronomy

D. Integrated topics (multidisciplinary)

1. Models
2. Equilibrium
3. Change
4. Evolution
5. Growth
6. Time/space
7. Systems
8. Cycles
9. Probability

II. Processes

A. Process/methods

1. Models
2. Assumptions
3. Communications
4. Measurement
5. Classification
6. Observation
7. Experimentation
8. Interpretation of data

B. Science and societal problems

1. Health and safety
2. Environment
3. Growth and resource management

C. Science and self

D. Science and technology (applied science)

1. Biological
2. Physical

E. Decision making

AFFECTIVE DOMAIN

A. Attitudes toward science classes

- To what extent are science classes enjoyable?
- To what extent does the student perceive individualization in science classes?
- To what extent do science teachers enjoy science and reflect that enjoyment to the students?
- Are science classes useful?
- What extracurricular science-related activities do the students pursue?

B. Vocational and educational intentions

- To what extent do students consider science as an area of further study and career possibilities?
- How do students rate a science-related vocation?

C. Personal involvement

- Do students recognize serious problems in the world today?
- Can students effectively do anything to solve major problems?
- Are students willing to help solve major problems?
- How often do students participate in activities that aid in solving major problems?

D. Tools--attributes

- Are the concepts and principles learned in science classes useful or applicable in everyday scientific investigation and decision making? In problem solving?

E. Confidence in science

- Determine the attitude students have toward the conduct and support (i.e., financial) for applied research, basic research

F. Controversial issues

- What are students' opinions and attitudes about allowing research in areas with potential hazards and risks?

G. Awareness

- Are students aware of the scientific process and the empirical nature of science?
- Are students aware of the tentativeness of scientific theories?

H. Experience

1. Experience--done something
   - Have students ever done science-related activities?
   - Would students like to do science-related activities?
2. Experience--seen something
   - Have students ever seen different events or activities related to science?
   - Would students like to see different events or activities related to science?
3. Experience--used something
   - Have students ever used various science-related objects?
   - Would students like to use various science-related objects?

4. Experience--visited a place

   - Have students ever visited various places related to science?
   - Would students like to visit various places related to science?

5. Experience--done experiments

   - Have students ever done experiments or activities with various science-related things?
   - Would students like to do experiments or activities with various science-related things?

MUSIC

I. Value music important as an important realm of human experience

   A. Be affectively responsive to music

   B. Be acquainted with music from different nations, cultures, periods, genres, and ethnic groups

   C. Value music in the life of the individual, family, and community

   D. Make and support aesthetic judgments about music

II. Perform music

   A. Sing (without score)

   B. Play (without score)

   C. Sing or play from a written score

   D. Sing or play a previously prepared piece

III. Create music

   A. Improvise

   B. Represent music symbolically

      1. Arrange

      2. Compose

IV. Identify the elements and expressive controls of music

   A. Identify the elements of music

      1. Rhythmic organization

      2. Pitch organization

      3. Tone quality

   B. Identify the relationships of elements in a given composition

   C. Demonstrate an understanding of a variety of musical terms, expression markings, and conducting gestures in a musical context

V. Identify and classify music historically and culturally

   A. Identify and describe the features that characterize a variety of folk, ethnic, popular, and art music

   B. Identify and describe the music and musical style of the various stylistic periods in Western civilization (e.g., medieval, renaissance, baroque, classical, romantic). Identify representative composers of each period

   C. Cite examples of ways in which man utilizes music in his social and cultural life

ART

I. Perceive and respond to aspects of art

   A. Recognize and describe the subject matter elements of works of art

   B. Go beyond the recognition of subject matter to the perception and description of formal qualities and expressive content (the combined effect of the subject matter and the specific visual form that characterizes a particular work of art)

II. Value art as an important realm of human experience

   A. Be affectively oriented toward art

   B. Participate in activities related to art

   C. Express reasonably sophisticated conceptions about and positive attitudes toward art and artists

III. Produce works of art

   A. Produce original and imaginative works of art

   B. Express visual ideas fluently

   C. Produce works of art with a particular composition, subject matter, expressive character, or expressive content

   D. Produce works of art that contain various visual conceptions

   E. Demonstrate knowledge and application of media, tools, techniques, and forming processes

IV. Know about art

   A. Recognize major figures and works in the history of art and understand their significance. (Significance as it is used here refers to such things as works of art that began new styles, markedly influenced subsequent works, changed the direction of art, contained visual and technical discoveries, expressed particularly well the spirit of their age, or are considered to be the major works of major artists.)

   B. Recognize styles of art, understand the concept of style, and analyze works of art on the basis of style

C. Know the history of man's art activity and understand the relation of one style or period to other styles and periods

D. Distinguish between factors of a work of art that relate principally to the personal style of the artist and factors that relate to the stylistic period or the entire age

E. Know and recognize the relationships that exist between art and the other disciplines of the humanities (literature, music, and particularly the history of ideas and philosophy) during a given period

V. Make and justify judgments about the aesthetic merit and quality of works of art

A. Make and justify judgments about aesthetic merit

B. Make and justify judgments about aesthetic quality

C. Apply specific criteria in judging works of art

D. Know and understand criteria for making aesthetic judgments

CAREER AND OCCUPATIONAL DEVELOPMENT

I. Knowledge, abilities, and attitudes relevant to career decisions

A. Awareness and knowledge of individual characteristics

1. Abilities

2. Interests

3. Values

B. Knowledge of career and occupational characteristics

1. Major duties

2. Entry requirements

3. Work conditions

4. Benefits and employment practices

5. Social and technological change

6. Occupational classification

C. Making and implementing career and occupational decisions

1. Individual characteristics and occupational requirements

2. Career decision making

3. Career preparation

4. Career modification or change

5. Sources of additional knowledge

II. Knowledge, abilities, and attitudes necessary for success in a career or occupation

A. Skills generally useful in careers

1. Numerical skills

2. Communication skills

3. Manual/perceptual skills

4. Information-processing, problem-solving, and decision-making skills

5. Interpersonal skills

6. Employment-seeking skills

7. Career-improvement skills

B. Personal characteristics related to career success

1. Responsibility and initiative

2. Adaptability to variable conditions

3. Attitudes and values

4. Personal fulfillment

- Citizenship and Social Studies Objectives: 1981-82 Assessment (1980)
- Science Objectives: Third Assessment (1979)
- Music Objectives: Second Assessment (1980)
- Art Objectives (1971)
- Career and Occupational Development Objectives: Second Assessment (1977)

The outlines in Table 1 encompass the NAEP objectives, and they represent--in fact, in most cases, they are--the questions that the assessments attempt to answer about performance in each learning area. In several areas, the booklets also contain further detail on objectives or questions within the finest level of the outline shown. Thus in science, in career and occupational development (COD), and social studies-citizenship, there is a great amount of additional detail. In science, this takes the form of "sample" objectives that are specifically said not to be definitive as far as the potential assessment exercises are concerned. COD objectives are stated in a quasi-behavioral form and in extensive detail, for example "Understand one's own abilities relative to those of others," and "Know where to find information regarding job openings." Each objective is further described as to its appropriateness and the sophistication presumed of persons at each age level. In writing, there is also additional detail by way of examples of the kinds of writing that might be done by persons at each of the age levels, but there is very little specificity as to how ability to write is really to be evaluated or scored at each age level.

The NAEP objectives are obviously not specific enough to uniquely define particular exercises as the appropriate measure of performance on the objectives as a whole. As we will elaborate later, they lack specification of the stimulus condition to be presented to the student or the criterion by which mastery is to be evaluated. NAEP exercises can be referenced to an objective they are supposed to assess, but many exercises with quite different skill requirements (and hence different performance outcomes) can be and have been developed and referenced to the same objec-

tive. NAEP exercises are not criterion-referenced. They can be said to be objectives-referenced. However, the converse inference, namely that the level of performance on a given exercise can be interpreted as indicative of the level of performance on the objective as a whole, is generally not warranted, even when the most specific NAEP objectives at the most detailed level are considered. This inference, of course, has been frequently made, and we will have more to say on that subject in the section on exercise development.

When the various sets of assessment objectives are arrayed in proximity, as in Table 1, it is difficult to avoid being struck by their very considerable differences. Objectives in the various areas are not organized by any consistent framework, and they are phrased very differently (e.g., as questions, as quasi-behavioral statements, or as topics). The objectives also differ widely in specificity from area to area and in how they treat cognitive versus affective objectives (i.e., whether affective objectives are integrated into the substantive material (e.g., in music, art, and social studies-citizenship) or kept as separate objectives (e.g., in science and math). The areas also differ in whether and how they recognize "process" objectives or objectives at differing levels of Bloom's taxonomy (e.g., the differences between recall or recognition of facts or principles in the subject area and the ability to apply these principles to the solution of problems or the evaluation of new situations). It is also evident that there is overlap between the areas. Applied mathematics objectives occur among the COD objectives, for example, and problem solving and thinking/information utilization skills are included in several areas. For the most part, these are linked to the subject matter of the area in question, but in COD, for example, they are treated as more general skills of use in all sorts of employment settings.

Finally, there are types of objectives found in each list that differ in terms of the aspect of the student to which they refer. Since they are not treated consistently from list to list, their similarities are harder to discern. Thus, objectives related to the individuals' capacity (what they can do when asked and appropriately motivated) are not readily discerned from objectives that relate to habitual response patterns (that is,

what the individual commonly does outside the assessment situation, for example in leisure activities), or from those that relate to what the person's intentions or dispositions might be with respect to future choices or in hypothetical situations.

An issue worth addressing at this point in NAEP's history is whether a conceptual framework can be developed to integrate and organize the several assessment areas. That this is feasible without doing violence to the separate disciplines has been demonstrated in other objectives-based assessments dealing with the same content areas and with all grades, K-12 (e.g., the PLAN Master Objectives, Westinghouse Learning Corporation, 1971). That this would have advantages in terms of communication with NAEP's audiences about NAEP objectives is fairly obvious. What is perhaps less obvious is that such a conceptualization could suggest other ways of assessing objectives or of clustering objectives and exercises for reporting purposes. Some consistent organization could also help identify any types of objectives that have been emphasized or omitted from particular assessment areas and could help suggest the most efficient and effective ways of measuring performance on certain objectives.

As noted above, the precise nature of the objectives in each assessment area has been determined primarily by the subject matter experts, educators, and lay public, with the selection of these persons, the timing of their work, and the process by which their input is secured being determined, at least at the policy level, by the Assessment Policy Committee (APC). The APC also sets the schedule and magnitude of the assessments in various areas. In reviewing the current NAEP objectives, we could only identify two areas, science and math, where there was an explicit plan for a differential emphasis between subareas or sets of objectives within a subject area. This differential emphasis is reported in the form of a matrix that indicates the percentage of exercises by age and content (e.g., number and numeration 40%, 40%, and 35% at ages 9, 13, and 17, respectively). In areas where the objectives have been either substantially revised or combined across two previously distinct assessment areas, it seems likely that the emphases among objectives have also changed over time and been reflected in differential numbers of exercises refer-

enced to particular objectives. It is not clear whether there have been formal, informal, or no priorities for assessment of particular objectives in most of the assessment areas.

We have attempted to induce the relative priorities among content areas by examining the NAEP assessment schedule, the numbers of released exercises in each area, and an estimate of the total number of exercises in each area. Table 2 shows the schedule of assessments by year and assessment area. Reading-literature (including the planned 1984-85 assessment), mathematics, science, and writing (including the upcoming assessment in 1983-84) have been most frequently assessed, with social studies-citizenship, art, and music next in frequency. Career and occupational development has only been assessed once, perhaps due to the cost of administration of the exercises originally developed in this area.

Table 3 provides another perspective on priorities. Assuming that roughly half of the objectives in each assessment have been released and that the repeat of materials in successive assessments is at roughly the same rate in all areas, one can estimate the average number of exercises per area per assessment cycle and from this obtain the percentage of all NAEP exercises in a given area in a full assessment cycle (i.e., after assessing each area once). These numbers are undoubtedly imprecise due to the unknown amount of duplication in exercises across assessments, but they generally accord with the priorities evident from the assessment frequency. However, social studies-citizenship appears to have relatively more exercises per assessment than its priority for assessment frequency would suggest, and writing, which has been frequently assessed, has relied upon fewer different exercises. The latter is presumably due to the fact that the writing exercises require production and are primarily in open-ended rather than multiple-choice format. They also are the longest exercises, requiring up to 20 minutes, whereas the average is about one minute (a range of about 30-90 seconds) for other items. Thus, if one were to look at priorities in terms of time devoted to the gathering of data in each area, the results would probably be more in line with the priorities inferable from the relative frequency of assessments in each area.

Table 2

Summary of NAEP Assessment Schedule

Year	Assessment Area								
	Reading/ Literature	Writing	Math	Citizenship/ Social Studies	Science	Career and Occupational Development	Art	Music	Special
1969-70		W		C	S				
1970-71	R,L								
1971-72				SS				Mu	
1972-73			M		S				
1973-74		W				COD			
1974-75	R						A		
1975-76			M	C/SS					
1976-77					S				+
1977-78			M						++
1978-79		W					A	Mu	
1979-80	R/L								
1980-81									
1981-82			M	C/SS	S				
1982-83									
1983-84*		W							
1984-85*	R/L								

* Planned

\+ Basic life skills (17-year-olds only), health and energy (adults only), reading and science (adults only)

++ Consumer skills (17-year-olds only)

Table 3

Summary of Released Exercises and Estimates and Relative Numbers of Exercises per Assessment by Assessment Area*

Area	(Number of Assessments)	Number of Exercises Released	Area as % of Total Released Exercises	Average Number of Exercises per Assessment**	% of Exercises in a Full Assessment Cycle	% of Exercises in Multiple Choice Format
Art	(2)	112	5%	112	6%	93%
Social Studies/ Citizenship	(2)	344	16%	344	19%	58%
Career and Occupa-pational Development	(1)	61	3%	122	7%	21%
Math	(3)***	494	23%	329	18%	56%
Music	(2)	155	7%	155	9%	78%
Reading, Literature	(2)***	275	13%	275	15%	79%
Science	(3)***	599	28%	399	22%	93%
Writing	(3)	90	4%	60	3%	17%
		2,130	99%	1,796	99%	
Special Problems		306				96%
		2,436				

* Data taken from Summary Table of National Assessment Released Exercises (10/81), enclosure in NAEP publication SY-01-36.

** Assumes that approximately 50% of exercises used have been released. Estimate equals 2 x (number released) divided by number of assessments; it does not take account of overlap in exercises between successive assessments, which would reduce the resultant estimates.

*** Summary table of released exercises does not indicate any release of exercises for Reading 1974-75, Math 1981-82, or Science 1981-82, hence the number of additional items used in these assessments was not determinable.

BEST COPY AVAILABLE

Future NAEP Assessment Objectives

The current NAEP assessment objectives in each area are the primary embodiment of the many highly specific questions that the assessment has undertaken to answer. The objectives (and exercises) may be aggregated in various patterns to answer a variety of additional questions, but it is clear that there must be considerable constancy in the basic objectives if progress is to be measured meaningfully over time. On the other hand, periodic review and revision of the objectives is needed to ensure that they remain relevant, and the discussion above has pointed out several aspects of the objectives that could be substantially improved, namely their consistency across areas, their specificity, and the matter of explicit priorities for inclusion of items in each area when it is assessed.

The Assessment Policy Committee is responsible for setting priorities for NAEP assessment, which includes responsibility for approving the procedures used to review and revise objectives, ratification of the resultant objectives for use in a particular assessment, and establishment of the assessment schedule. Normally, review of the objectives takes place prior to their use in an assessment in that area. However, we recommend that in the future the APC provide guidelines to the subject matter consultants and to the Area Advisory Committees so as to better integrate their efforts and arrive at a more consistent structure and format for the objectives across areas. We also recommend that specific efforts be undertaken to provide empirical data on educational objectives to aid the APC in its deliberations, especially in the matter of setting priorities. Finally, with the concurrence of the APC, the NAEP contractor should provide a more formal structure to the process by which the input and consensus of a broad sampling of educational and lay constituencies is obtained concerning assessment objectives, accomplishing this without requiring the time and expense of a process composed primarily of committee deliberations.

Sample issues. The following is a brief discussion of some of the important issues that we see in conjunction with NAEP objectives.

1. The need for a conceptual framework to relate the various assessment areas and objectives to each other, to the broader goals of the assessment, and to the plans for analysis and reporting.

NAEP is in need of a clearly articulated conceptual framework that would describe the general dimensions on which it is important to assess educational progress and relate these, in progressively more specific hierarchical fashion, to the assessment of major subject areas and to the objectives within each area. The framework should also make it clear how other information gathered by NAEP is to be used to assess educational progress. That is, it should specify what analyses will be done to compare performance between various populations or areas. And finally, the framework should clarify what aspects of education and educational progress are currently adequately addressed by data from other sources (e.g., NCES data on educational resources) and for what aspects there is no consistent, adequate data base. While the originators of NAEP may have felt that they had such a framework in mind, it has not been explicit and accessible. Its absence is part of the reason that NAEP does not communicate well to any audience. As times change, as other data bases develop, and as the needs for particular types of information become clearer (e.g., on courses offered and course enrollment), the absence of an explicit conceptual framework for NAEP makes it difficult to define or redefine the role of a national assessment appropriately. We propose that the time of transition is an appropriate, if not long overdue, time for the APC to consider this issue.

One general conceptual framework that might be considered as an option for NAEP would be to begin with two very general purposes of education: (1) preparing students at each grade level to pursue and profit from subsequent education at the next level and (2) preparing students for productive and fulfilling roles in society. NAEP's purpose, then, might be expressed as that of assessing the knowledge, skills, understanding, and attitudes of students at various grade levels in relation to the requirements for attaining these goals. From a framework then would naturally follow the criteria against which objectives might be reviewed at each grade level, namely, the extent to which students' progress in the

next grade level or phase of their education would be impaired by a failure to master a given objective, and the extent to which their productivity and fulfillment in social roles outside of school would be impaired by a failure to master the objective. Such a framework would also provide a focus for integrating empirical evidence, not just opinion, into the specification of objectives. Empirical data exist on the predictive relationships between prior knowledge, skills, and understanding and subsequent academic progress. There is evidence, albeit less extensive, on the knowledge and skill requirements of various adult roles.

This is hardly a new concept. As early as 1950, John Flanagan and others were proposing the empirical definition of "critical requirements" as the primary basis for extablishing educational goals and assessing progress (e.g., Flanagan, 1950). This framework would also suggest a dual focus for reporting of assessment results--on the one hand in relation to students' preparation for subsequent education and on the other hand in relation to the performance requirements of various adult roles, for example work, citizenship, health, as a consumer, and as a spouse and parent. A number of educators concerned about adult knowledge have conceptualized the general areas in which knowledge is applied in ways that may be of use to NAEP (e.g., the Northcutt study conducted at the University of Texas, 1975). An alternative conceptualization of the relevant areas of adult life for which specific knowledge, skills, attitudes, and so on are required--areas for which schools attempt (in varying degrees) to prepare students and in terms of which NAEP objectives and exercises might be evaluated and results from NAEP reported--are the five critical areas and 15 component dimensions or factors related to the quality of life of adults as developed by Flanagan and Russ-Eft (1975) and Flanagan (1978) and shown in Table 4.

Obviously, different schools and different "schools of thought" tend to place more responsibility on preparation related to some areas than to others. The struggle to arrive at a set of general goals for education or even a framework for stating goals has been a long and arduous one, as Flanagan has pointed out (1978). But considerable progress has been made, notably in the years since NAEP began and in some measure aided by the

Table 4

Critical Factors Related to the Quality of Life of Adults

PHYSICAL AND MENTAL HEALTH DEVELOPMENT

A. Health and personal safety (98%)*

Enjoying freedom from sickness, possessing physical and mental fitness, avoiding accidents and other health hazards. Problems related to alcohol, drugs, death, and aging are also included. Effective treatment of health problems is a large component.

B. Personal understanding and planning (88%)

Developing and gaining orientation, purpose, and guiding principles for one's life. This may involve becoming more mature, gaining insight into and acceptance of one's assets and limitations, experiencing and awareness of personal growth and development, and realizing the ability to influence the course of one's life significantly. It also includes making decisions and planning life activities and roles. For some people, a major component arises from religious or spiritual experiences or activities.

INTERPERSONAL DEVELOPMENT

C. Relations with parents, siblings, or other relatives (76%)

Having parents, siblings, or other relatives. In these relationships, one experiences communicating with or doing things with them, visiting, enjoying, sharing, understanding, being helped by, and helping them. The feeling of belonging and having someone to discuss things with is a large component.

D. Relations with friends (75%)

Having close friends. In these relationships one shares activities, interests, and views. Important aspects of these relationships involve being accepted, visiting, giving and receiving help, love, trust, support, and guidance.

E. Relations with spouse (girlfriend or boyfriend) (92%)

Being married or having a girlfriend or boyfriend. The relationship involves love, companionship, sexual satisfaction, understanding, communication, appreciation, devotion, and contentment.

F. Socializing (51%)

Entertaining at home or elsewhere, attending parties or other social gatherings, meeting new participation in socializing organizations and clubs.

G. Having and raising children (88%)

Having children and becoming a parent. This relationship involves watching their development, spending time with them and enjoying them. Also included are things like molding, guiding, helping, appreciating, and learning from them and with them.

INTELLECTUAL AND CREATIVE DEVELOPMENT

H. Intellectual development (84%)

Learning, attending school, acquiring desired knowledge and mental abilities, graduating, and problem solving. Other aspects involve improving understanding, comprehension or appreciation in an intellectual area through activities in or out of school.

I. Creativity and personal expression (50%)

Showing ingenuity, originality, imagination in music, art, writing, handicrafts, drama, photography, practical or scientific matters, or everyday activities. This also includes expressing oneself through a collection, a personal project, or an accomplishment or achievement.

J. Passive and observational recreational activities (55%)

Participating in various kinds of passive recreation, such as watching television, listening to music, reading, going to the movies, and going to entertainment or sports events. It also involves appreciating the art and beauty in many aspects of life.

K. Active and participatory recreational activities (53%)

Participating in various kinds of active recreation, such as sports, hunting, fishing, boating, camping, vacation travel, and sightseeing. This may also involve such activities as playing sedentary or active games, singing, playing an instrument, dancing, or acting.

CAREER DEVELOPMENT

L. Occupational role (job) (90%)

Having interesting, challenging, rewarding, worthwhile work in a job or home. This includes doing well, using one's abilities, learning and producing, obtaining recognition, and accomplishing on the job.

M. Material well-being and financial security (73%)

Having good food, home, possessions, comforts, and expectations of these for the future. Money and financial security are typically important factors. For most people filling these needs is primarily related to their efforts or those of their spouse.

CIVIC DEVELOPMENT

N. Activities relating to local and national governments (43%)

Keeping informed through the media; participating by voting and other communications; having and appreciating one's political, social, and religious freedom. One component of this includes having living conditions affected by regulations, laws, procedures, and policies of governing agencies and the individuals and groups that influence and operate them.

O. Activities related to helping or encouraging other people (63%)

Helping or encouraging adults or children (other than relatives or close friends). This can be done through one's efforts as an individual or as a member of some organization, such as a church, club, or volunteer group, that works for the benefit of other people.

Note: From Table 2 of Flanagan and Russ-Eft (1975).

*The numbers in parentheses are the percentages of 30-year-olds who considered the component to be important or very important for their quality of life (Flanagan & Russ-Eft, 1975).

BEST COPY AVAILABLE

NAEP effort. Table 5, taken from Flanagan (1978), summarizes the broad-gauged goals of elementary and secondary eduction as enunciated by state governments within the past decade. Flanagan has noted their similarity to the critical dimensions affecting the quality of life of adults that have emerged from his research, as well as their similarity to the cardinal principles of secondary education as set forth in 1918 by the Commission on the Reorganization of Secondary Education--"health, command of fundamental processes, worthy home membership, vocation, civic education, worthy use of leisure, and ethical character."

The value of having a concrete list of general educational goals and critical areas of adult life is that it can help clarify the various perspectives in terms of which the assessment objectives might be evaluated. In fact, such a concrete description can help various constituencies to clarify the perspectives from which they are evaluating assessment objectives, highlighting similarities and differences relative to the views of others, and aiding in reaching a consensus. The convergence in the above descriptions suggests that the prospects are good for achieving a consensus on a conceptual framework for NAEP.

To this point in time, NAEP has not had any general conceptual framework that is linked to the overall goals of education. Only one of a number of alternative possibilities is outlined above, albeit one that has much to recommend it. The potential benefits of adopting a workable framework are immense, however, and we recommend that alternative options be identified and presented to the APC to begin consideration of this issue.

2. <u>The development of priorities for assessment among areas, subareas, and specific objectives</u>.

This issue is perhaps one of the most sensitive for NAEP and is central to the responsibilities of the APC. Up to now, priorities have been primarily reflected in the frequency of the assessments in various areas, as noted earlier. One could undertake to count the numbers of items in each assessment that have been related to each objective and thus

Table 5

Educational Goals for Elementary and Secondary Education as Adopted by State Governments

Physical and Material Well Being	*Personal Development and Fulfillment*
22 A. Each individual must *develop an understanding* of the principles involved in the production of goods and services and *of the skills relating to the management of personal resources.*	47 H. Each individual must *master the basic skills* of reading, writing, speaking, listening, computation, and problem solving.
41 B. Each individual must *acquire* good health and safety habits and an *understanding of the conditions necessary for physical and mental well-being.*	38 I. Each individual must *master the skills of constructive and critical thinking and decision making* so that he or she can deal effectively with problems in an open-minded and adaptable manner.
29 C. Each individual must *develop the knowledge and respect necessary for the maintenance, appreciation, protection, and improvement of the physical environment.*	36 J. Each individual must *gain knowledge* of the human achievement and experience in the areas of *natural sciences, social sciences, humanities, creative and fine arts.*
Relations with Other People	41 K. Each individual must *gain an eagerness for learning* and self-development beyond the formal schooling process.
24 D. Each individual must learn the rights and responsibilities of family members and *prepare for family life.*	40 L. Each individual must *develop a positive self-image* and an understanding and appreciation of his or her unique capacities, interests, and goals.
36 E. Each individual must learn to *develop and maintain interpersonal relationships* and have command of social skills.	45 M. Each individual must select and *prepare for a career* of his or her choice consistent with his or her capabilities, aptitudes, desires, and the needs of society.
Social, Community, and Civic Activities	35 N. Each individual must *develop* a personal philosophy and *a basic set of values, morals, and ethics* acceptable to society.
39 F. Each individual must come to *understand and appreciate different cultures,* governments, races, generations, and life styles.	29 O. Each individual must *acquire the desire and ability to express himself or herself creatively* and to appreciate creativity in others.
43 G. Each individual must *learn the rights and responsibilities of citizens* of the community, state, and nation.	Recreation
	27 P. Each individual must have knowledge of and *skills in recreation and leisure-time activities* for nonvocational use of time.

Notes: Information in table is from state governments except Arkansas, Indiana, and Minnesota.

The figure at the left of each goal indicates the number of states that have adopted it as one of their educational goals.

infer the implicit priorities that NAEP has observed. However, no explicit priorities are evident except for proportional emphasis (percent of exercises) in the subareas of math and of science. The assumption appears to have been that all objectives are important, if not equally important. The problem is that the numbers of objectives in many areas are more numerous than the number of exercises that could be administered in a given assessment, hence some objectives have obviously not been assessed, at least not consistently. Instructions for NAEP exercise development do not appear to set priorities, and the primary considerations in deciding on the inclusion of exercises are the following:

- Whether the exercise is referenceable to any one of the NAEP objectives--i.e., appears to be a legitimate assessment of the objective.
- Whether the exercise is relatively low in cost to administer--individually administered exercises are more costly and have been dropped in recent years; performance exercises (e.g., measurement of small group participation in citizenship) are also more costly.

This has inevitably meant that some objectives have had lower de facto priority--either because relevant items were not generated by exercise developers in the first place or because no cost-feasible means of testing them had been devised. Such limits are inevitable--resources are never unlimited, and choices have to be made. NAEP could take a major step forward in communication with its audience and in stabilizing the indices of education progress that it can provide, however, by developing explicit priorities to guide the development of exercises and their assembly into assessment packages.

We recommend that this issue be raised with the APC. Because of the sensitivity of this matter and the diversity of perspectives that potentially exist on priorities, we also suggest here a multifaceted approach to providing NAEP with a defensible method of priority setting, one that makes the most appropriate use of the judgment and experience of all sorts of lay and professional persons concerned with education. This approach has been used very effectively in the past to define the objectives of an

assessment program. This past experience is useful to recount in that it indicates the potential of the approach to provide a reliable basis for determining and stabilizing the knowledge to be assessed and its potential for clarifying and reconciling what might appear to be irreconcilable differences in the perspectives among those who will be affected by the assessment. The circumstances of the previous use of this approach might appear different from the context of NAEP, and in fact the purpose was quite different--the development of content specifications for science tests to evaluate applicants to medical school. However, many of the basic issues and problems in arriving at a consensus on what was to be tested were similar and the lessons learned appear applicable.

For many years before AIR undertook revision of the admissions test given by medical schools (the MCAT), there had been increasing debate and dissatisfaction with the previous assessment tool. Meetings were held over a two-year period concerning the problem of how to evaluate applicants. A feeling had arisen that the scientific preparation and academic quality of the applicants was such that attempting to make fine distinctions on the basis of more and more advanced and specialized preparation in science was ill-advised. Conversely, more nontraditional students, women and members of minority groups, were applying, and there was a need, not only to provide a fair assessment of these applicants, but to provide a more diagnostic assessment of their strengths and any needs for remediation were they to be admitted. Yet other traditional differences in perspective existed between those teaching the basic sciences in the first two years of medical school, the clinical science faculty, and practicing physicians, to say nothing of the perspectives of undergraduate college faculty and advisors. College faculty had serious concerns about the impact of a national test on their curriculum, which was not exclusively aimed at preparation for professional school.

When AIR was placed under contract to develop the new assessment specifications in 1974, it was in part because we offered a means of resolving these differences. Among the persons most likely to be affected, however, there was still a high level of anxiety and cynicism about the possibility of achieving a consensus.

AIR consulted with educators and studied curriculum outlines and texts used widely in introductory biology, general chemistry, organic chemistry, and physics at all sorts of colleges across the country, including both the most and least selective schools and those with various missions and student body compositions. The result was a comprehensive and detailed outline of topics taught at any of these schools in the four subjects. The deans of all U.S. medical schools as well as representatives of all the special interest groups were then asked to provide nominations of persons qualified by experience and breadth of perspective to evaluate the relevance of prior scientific knowledge to the medical school curriculum and to medical practice from each of the following groups: basic science faculty, clinical science faculty, physicians in practice, and senior medical students and residents. We specially sought nominations of women and minority group members. In the end, over 300 highly qualified nominees from all of the concerned groups and segments of the medical education community were constituted as an evaluation panel. The panel did not meet. Each individual made deletions, additions, or modifications as needed and then provided independent evaluations of each topic on the detailed outline. The first rating was of the extent to which a student's mastery of the curriculum in medical school would be impaired by a lack of understanding of the topic at the time of entrance, from "not at all impaired" to "seriously impaired." The second evaluation was of the frequency with which the topic was utilized in the practice of medicine. It was evident that the ratings were done in a thorough and thoughtful manner.

When the results were examined, it became clear that there was a high degree of agreement among raters, regardless of their position or personal background, on the importance of different topics relative to each other. The differences showed up primarily in the overall levels of the ratings, which did differ between groups. Thus, students tended to see the science material, as a whole, as being less important to mastery of the curriculum, as did women and members of minority groups; medical basic science faculty gave it highest overall importance. The agreement on relative importance of topics, however, provided the key to setting priorities for the inclusion of items on various topics in the test, by providing a basis for

eliminating material the understanding of which was judged unlikely to affect progress, either because it was not relevant to medical training or practice or because it was covered in medical school and was not pre-requisite.

What is proposed for NAEP is that the process of review and revision of objectives be structured and supplemented by a similar systematic survey to obtain the judgments of relevant knowledgeable persons on the importance of each NAEP objective and possible additional objectives in terms of the following criteria:

- the extent to which a student's mastery of the curriculum at the next grade level would be impaired if he or she had not mastered the objective;
- the extent to which the knowledge skills, understanding, or attitudes embodied in the objective make a difference in adult daily living.

The individuals surveyed should include concerned individuals with recognized breadth of perspective and judgment representing all facets of the educational world, including:

chief state school officers
legislators
state and local school board members
school district superintendents
school administrators
teacher educators
classroom teachers
students
lay citizens
curriculum specialists
subject matter specialists
representatives of major professions and career fields

These individuals should be selected for their breadth of expertise and experience and their interest in education. They should also be selected so as to include persons from all regions of the country, diverse ethnic backgrounds, and both sexes. A broad sample of organizations and agencies representing the perspectives to be included should be enlisted to nominate

persons to be surveyed. Thus, for example, nominations could be invited from the Council of Chief State School Officers, the National Association of State Boards of Education, the American Association of School Administrators, the National School Board Association, the National Parent-Teacher Association, the National Education Association, the American Federation of Teachers, and so on. A list of relevant organizations and individuals to be invited to submit nominations should be drawn up and specifically reviewed by the APC. The goal should be broad inclusiveness, with a view to screening only for expertise and relevant experience and to ensuring sufficient representation of persons from various perspectives and backgrounds. An Objectives Evaluation Panel consisting of about 300-400 persons is necessary to ensure sufficient numbers from each subgroup to permit meaningful comparison of ratings among subgroups defined in terms of their position within the world of education and the background of the rater.

We recommend that the process of selection of the Objectives Evaluation Panel also be used to identify a subset of subject matter experts, educators, and lay persons with appropriate expertise and interest to serve on eight Assessment Area Advisory Committees to the separate assessment areas.

Careful thought will need to be given to the precise task to be presented to the Panel. As they are currently written, the NAEP objectives are too variable in format and specificity to be rated directly. Therefore, prior to the survey, NAEP staff, with the assistance of expert consultants in each area, would have to prepare a master list of educational objectives in consistent form for each assessment area. This list would include all current NAEP objectives plus any additional objectives suggested by NAEP consultants and staff and by consideration of sets of objectives developed by other groups. Because the entire set of detailed objectives is likely to be extremely large, it should be organized hierarchically within subject matter areas and grade levels: primary, intermediate, and secondary.

The overall goal is to obtain evaluations in the context of the entire set of NAEP objectives, not only on a discipline-by-discipline basis. Thus, all panelists should be asked to rate all objectives on the two criteria listed above. Specific subsets of the panel, notably subject matter specialists, teachers, and those dealing with curriculum in particular subject areas and at particular grade levels, should be asked to make ratings at a more detailed level within their special areas of expertise. The detailed objectives should be available to all evaluators, and the evaluators should be asked to eliminate objectives they consider unimportant in each area prior to assigning their ratings.

The results of this survey would then be analyzed by NAEP staff both for the ratings of the panel as a whole and by subgroup, and the results should be summarized for the APC and the Assessment Area Advisory Committees. NAEP staff would use the results to prepare:

1. recommended revisions to the objectives,

2. a plan for emphasis in the assessments in terms of the numbers of items to be allotted to assessment of particular objectives, and

3. sets of overall exercise scoring weights to be used in constructing summary performance measures based upon the rated importance of the objectives.

Different weightings could, for example, reflect the importance of objectives in relation to the curriculum progress perspective (criterion 1) and in relation to the knowledge application perspectives (criterion 2). Other weightings might reflect the perspectives of particular regions of the country. Eventually, specific state and local education agencies and other bodies might wish to request special reports with weightings of objectives in relation to some alternative pattern of emphasis. As long as the pattern of development and administration of exercises across objectives is controlled appropriately, the number of special indices that may be reported is virtually unlimited. The ratings, however, would provide a very defensible basis for the development of basic indices of progress to be reported on a regular annual basis. And given some stan-

dardization of the format of objectives across areas, it would be easy to generate special across-area indices such as indices related to particular areas of adult life or to particular skills (e.g., decision-making).

In future years, the rating process could be repeated in briefer format to ensure that the objectives remain current or are revised accordingly. Our experience suggests that this sort of process yields results that are much more stable and reliable over time than the results of committee consensus-development efforts. Thus, we anticipate that it would be desirable to check on the currency of NAEP's objectives at 4-5 year intervals, with appreciable shifts anticipated on a longer time scale, perhaps over a decade.

3. The development of criteria by which student performance is to be evaluated.

From its inception, NAEP has wisely avoided attempting to set national "performance standards" against which schools, districts, and states might be compared--whether these be so-called minimum standards or standards of excellence. There is an issue, however, which is often confused with standard-setting, that does need to be considered in connection with the statement of individual objectives for NAEP. The issue is whether an attempt should be made to bring NAEP objectives closer to the ideal of "true" educational objectives--that is, "instructional outcomes described in performance terms" (see the Foreword by William M. Shanner to the PLAN Master Objectives, 1971, reproduced here as Appendix A). As Shanner pointed out, even some of the systematically constructed objectives of the PLAN curriculum for grades K-12 in language arts, mathematics, science, and social studies and guidance were open to multiple interpretation because they lacked a description of the stimulus condition under which a student was to perform or were simply statements that failed to suggest any sort of criteria. NAEP objectives, drafted originally in a period when the concept of instructional objectives was less familiar and when the thought of any kind of national assessment was potentially threatening, are subject to similar criticism, but the problem is somewhat more per-

vasive. As Shanner has pointed out, "to have critical comments made about one's objectives should be taken as a compliment, since this can only happen when one has taken the trouble to think them out and write them down." In this sense, NAEP is deserving of high praise for its efforts to make explicit many important objectives of education. However, there are important reasons why objectives should specify stimulus conditions--e.g., "Given a written passage whose tone makes us judge a character's action unfavorably,..." or "Given a bank's interest rate, a credit union's interest rate, an amount of money to be borrowed,..." and so on. They should then specify what the student is to do and the criterion by which performance is to be judged adequate or inadequate--in the latter example, "determine whether a loan from the bank or one from the credit union would cost less." Another example might be "Given a list of staements describing group relationships, recognize those that show prejudice and those that do not," or "Given the major digestive structure of humans, identify the function of each structure," or again "Identify three ways in which an individual in the United States can influence the decisions made by his elected representatives."

Avoiding ambiguity and properly stating the detailed assessment objectives is the only way of providing effective guidance for exercise development, and in the end, the exercises, not the objectives themselves, provide the measures of educational progress. Instructing exercise developers in a loose fashion to "write items that are a direct measure of some knowledge, skill, or attitude stated in the objective" or to "be sure that they measure something which will be meaningful to report" does not provide them with enough guidance to ensure that the exercise will be a valid measure of the objective. The basic objectives should do much more to constrain the writer to an appropriate task to present to the student in order to measure performance on each objective.

Annual Assessments of Learning Areas

There is widespread agreement that the deterioration of performance of NAEP to the point that assessments cannot even be conducted once a year is unacceptable. Moreover, we claim, the early design decision to focus each year on a different assessment (or, originally, assessments) is not optimal for efficiently achieving the goals for which NAEP has been and is intended.

We recommend that NAEP carry out annual assessments, spanning the space of skills that make up educational progress each year. This change to NAEP will

(1) increase its utility by making it possible to respond to needs for data on emerging policy and research issues;

(2) increase its utility by increasing the power and stability of the educational progress time series;

(3) increase its utility by creating a basis for estimating the relations of educational achievement in different areas to each other as well as to program factors;

(4) increase its efficiency by eliminating redundancies across areas;

(5) increase its efficiency by introducing a smoother flow of exercise development, data collection, and analysis and reporting activities; and

(6) increase its acceptability to students and teachers by providing exercise packages that are more interesting and have higher face validity for the goals of assessment.

It may appear difficult to reconcile these numerous benefits with the fact that the direction NAEP has taken has been away from broad annual assessments. In this chapter, we will (1) present an overview of our recommended change, (2) explain the benefits listed above and how they can be achieved, and (3) describe and estimate the costs associated with this change.

General Specifications

Evaluation of a strategy such as this requires joint consideration of a variety of factors. To provide a basis for the evaluation of annual assessments, we describe here the overall design we have in mind. We recommend that the NAEP grantee carry out a single assessment each year, using matrix sampling as before, but covering objectives in most or all of the areas that NAEP was designed to address, including:

Reading	Other Areas:
Writing	Career Development
Mathematics	Foreign Language
Science	Art
Social Studies	Music
	Health

with exercises designed so as to assess:

Knowledge Acquisition
Internalization of Processes
Ability to Apply Skills
Attitudes toward Skills.

The number of exercises in each content area may change from year to year as information needs as well as the skills to be learned in schools change. The previous chapter discussed the process of objectives refinement.

We assume that between 500 and 700 exercises will be assessed for each age group each year, in roughly 30 packets of 20 exercises each. The critical change is that each packet would cover most, if not all, of the areas of the assessment. To allow estimation of correlations between items in different packets, each student would complete three of these packets. In carrying out this plan, each selected student would need to take part in a two-hour testing session. (We have not encountered any strong opposition to this increase in testing time in our discussions of this issue with leading state educators, especially if useful feedback

could be given to the schools; however, our plan would also work if testing could not be done in two-hour sessions.)

Under this design, somewhat fewer students are needed at each age group or grade than in past assessments (roughly 20% fewer), because of the doubled administration time and the increased power and efficiency provided by using integrated instrumentation each year. Each assessment exercise will be given to roughly 2000 students (requiring a total of 20,000 students in each age group), and every exercise will be paired with every other exercise for approximately 140 students. Although the data base for each exercise taken by itself is smaller than in some of the past assessments, the power of the assessment will be much greater because (1) exercises will be combined rationally into composite scores and (2) results of assessments can be combined, as well as compared, from each year to the next.

The schedule of activities for NAEP under this modified design would constitute a two-part "fugue." The time period for a particular assessment, from initiation of exercise writing to publication of the main report and public-use tapes, would be two years (one year of preparation and one of data collection), so that at any particular time work will be in progress for two successive assessments. At the beginning of each two-year assessment period, projections of issues that will be important two years later will be made and presented to the Assessment Policy Committee. These deliberations will lead to the weighting of different areas in the development of the assessment forms (i.e., how many items to include in each area) and possibly to the occasional identification of new objectives that may need to be added to the assessment. They may also lead to plans for oversampling of certain subpopulations of schools or students during the assessment.

Exercise writing can occur throughout the year at times convenient to the item writers, but during the first three months of each assessment period, the goal will be to complete the pool of exercises to be used in that assessment. The following six months will be used for review, try-

outs, revision, and approval of drafted forms. At the same time, schools will be selected and contacted for participation in the following year.

The instruments that are developed will be administered according to a schedule like that currently in place, in order to maximize the comparability of the data across time. Activities in the spring of each year will focus, rather intensely, on data processing and the production of reports and public-use tapes. In this way, the results of each annual assessment can be released in the summer following the assessment. Of course, this does not mean that numerous secondary analyses will not be carried out at a more deliberate pace.

Potential Benefits

Let us now consider the benefits that will accrue due to this innovation in design.

First, because all content areas are potentially covered by the assessment each year, the data base will be much more responsive to program evaluation and policy needs, both before data collection and after the fact. As soon as an issue appears to be emerging, while the specific objectives for one year's assessment are being weighted, planning can be carried out to collect the background and program data from the students and schools that will make the achivement data relevant to the information needs. There will be no need to wait several years for the next cycle in which the relevant assessment (e.g., science or reading) will be performed.

Even after the emphases of an assessment have been decided upon, the possibility for annual comparisons, related to program data collected from a different source, will dramatically increase the relevance of the achievement data to policy research and evaluation needs. For example, with the current design, it is impossible to evaluate the possible effects of the change from "Title I" to "Chapter I" on achievement, because the basic skills assessments have insufficient "resolution." If comparable assessments had been carried out each year between 1975 and 1985, we would

be in a better position to make statements about the importance, or lack thereof, of the federal role in compensatory education.

This design is also highly compatible with a partially longitudinal school sampling design, in which each school would be visited in two or more assessment years. The use of an integrated annual assessment instrument covering essentially the same areas each year is, of course, necessary if the longitudinal study is to be part of the overall assessment.

Even if there were no other advantages, this greater flexibility for use of the NAEP data to address policy research and evaluation questions would, we believe, outweigh the costs associated with the change (which are discussed in the section below).

The second major advantage that derives from annual assessments that integrate most or all content areas is the greater stability and power in the time series. Over 15 years, at most four assessments have been performed in each content area, and this frequency is quite insufficient to begin to develop time-series projections or tests of hypotheses about relations of various factors to educational achievement. At the present rate, twenty additional years will be required before the data series become useful for these purposes; however, addition of a new data point every year, even if it were somewhat less reliable, would dramatically reduce the waiting time before the data could be used for econometric and other modeling. Moreover, the currently accepted five-year periodicity for the core assessments is too long to provide the basis for investigating relations to events that change from year to year. If the math assessment results in 1982 differ from the results in 1978, it is impossible to relate the changes to exogenous factors that changed over those four years. If math assessments had been carried out annually, however, possible effects of such factors as the growth of computer awareness, of unemployment, or of school closures on math skills or attitudes might be examined.

The third way in which the proposed design would increase the utility of NAEP data is by providing data for estimating relations among achieve-

ment in different areas. Because each exercise set given to a student would span the range of educational objectives, and because the recommended matrix sampling design would ensure that an adequate number of students respond to each pair of exercises (by combining all possible pairs of packets of 20 exercises when composing the two-hour test booklets), complete interitem correlation matrices could be estimated. This will open up a broad vista of analytical uses for NAEP data not now possible.

Nearly all policy research issues, evaluation issues, or educational practice issues that can be addressed empirically require data on relations, not merely on levels of single variables. In many cases, the most important relations are between an achievement score and a vector of hypothesized factors that affect that score. This is not the only important type of relation, however. Relations among achievement test areas are also important. For example, the following issues require these relations.

- Is low science achievement by particular target groups related to low reading scores?
- Does a program that raises reading scores also raise science scores?
- Are problem solving skills generalized across the areas of science, math, social studies, and reading?
- What abilities are related to a positive attitude toward activities in science, or in music or art?
- For which content areas is variance between schools largest, for which is variance within schools largest?
- How many factors characterize within-school achievement? Between-school achievement?

Each of these three ways of increasing utility, greater responsiveness, greater power and stability, and greater research applicability would itself outweigh the costs of this change in the design of NAEP, but there are also increases in the efficiency of NAEP data collection that compound the advantages of the integrated annual assessment.

First, by considering the space of educational objectives in a unified framework, substantial redundancies in test items can be eliminated. For example, it will be unnecessary to include subject matter reading items twice, both as assessments of reading and as assessments of the subject matter. And, to the extent that problem solving involves the same skills across areas, this type of skill need not be covered as completely in each and every area. The implication of the elimination of these redundancies is that more independent information can be gathered in the same time period. Items whose response could be very well predicted from other reponses can be replaced with more informative items.

Another benefit to efficiency is in the flow of operations in the conduct of NAEP. "Start-up" costs for different assessments will be virtually eliminated because the assessments will be continuously on-going. In particular:

- the pool of exercise writers will be continuously monitored and improved in all subject matter areas;
- exercise writing can be done on a convenience schedule, which will increase the availability of top levels of expertise;
- forms preparation and clearance and procedures will be much more similar from year to year;
- requirements for data collection staff training will be reduced because of the similarity of procedures from year to year;
- psychometric analyses aimed at improving the item pool can be standardized, thus reducing cost; and
- a common format for an annual Report Card (an annual National Assessment summary report) will facilitate its quick production each year.

Finally, we expect that the new format, with items from a variety of areas, will be better received by both schools and students. Students will like participating better because the assessment will be both less boring and less intimidating. The reduction in intimidation will come from the fact that it will be clearer that the assessment is really not a

test of any individual student, because so little time is given to any particular content area. The attraction to schools will be that the assessment package in any particular year more clearly covers the breadth of the educational objectives of the school, so that if the school chose to extend the testing to a larger fraction of its students, the NAEP instrument would be a reasonable assessment tool for the school. The same logic also applies at the LEA and SEA levels. Consequently, we expect that the change in design may actually reduce efforts needed to induce schools and students to provide the required data, even with the proposed increase in individual testing time.

Potential Costs

On the cost side of the equation, there are two factors to be considered: (1) that data collection will occur each year, rather than every other year, and (2) that the assessment will span the range of subject matter areas each year, rather than focusing on one or two areas. Concerning the first point, AIR has estimated costs and finds that there is no reason to suppose that assessments cannot be carried out every year within the authorized budget. The more interesting questions center on the costs of an integrated assessment vs. separate subject matter assessments.

A first reaction to this proposal for change in the NAEP design might be that there is only one essential cost increase that must be taken into account in evaluating its feasibility and desirability: the cost of increased testing time per student, if we are to conduct two-hour sessions instead of one-hour sessions. However, in order not to overlook any substantial cost component, we must consider all three phases of the assessment in detail: instrument development, data collection, and analysis and reporting. The increased testing time, of course, falls in the category of data collection costs.

When comparing an integrated assessment design to separate assessments of content areas, the instruments would be made up of approximately the

same number of items, of approximately the same types, so that increases in instrument development costs would be of secondary importance at most. We have noted that, in fact, the integrated and continuous nature of the assessments may actually improve the efficiency of item development.

Two cost increments may result, however. First, a choice will arise that was not previously present: quantitative choices of how many items in each subject matter area to include in each assessment. In one sense, this is an easier decision than deciding whether or not to perform an assessment in an area like consumer education, for example, because a few items can be added or dropped much more inexpensively than whole assessments can be mounted. On the other hand, we recommend a careful analysis of issues and objectives at the beginning of each assessment and consideration by the Assessment Policy Committee (APC) of the appropriateness of the coverage proposed for each area.

This choice may require substantial discussion, at least in the first year or two, because there has heretofore not been an opportunity to make such choices between areas in the design of NAEP. While the barrier of mounting a separate assessment may seem to many to be an acceptable excuse for failing to include some subject matter area, that argument does not hold for addition of a dozen items. To avoid undue expenditure of effort in this consideration, the NAEP grantee will need to prepare justification for the choices prior to presentation to the APC, while being willing to alter the design on the basis of the deliberations of that committee. One such justification might be the percentage of school time allocated currently to the particular areas; another might be the relative number of items assessed by NAEP, summed over the past dozen years. More forward-looking criteria include focus on skills needed for careers, reducing emphasis on skills that are becoming obsolete, and emphasizing areas about which policy issues are arising.

The second potential cost increment involves the prudent review of the coverage of objectives in each area by professional associations that have some responsibility for curricula in those areas. Although in the past it was necessary to maintain close communications with a particular

association only when planning an assessment in that area, a more continuous interaction will be required when all areas are being simultaneously assessed. In spite of the fact that the resulting interactions will require careful management, we believe that they will ultimately benefit the effort to assess educational progress as a whole.

The major cost increment, as noted above, is in the data collection phase. This is primarily a cost to the participating schools and their students, rather than a direct cost to the government, so its analysis is particularly critical.

Rather than conduct an assessment involving 25,000 students for one hour each, our proposal is to conduct an assessment involving 20,000 students for two hours each. The reasons that two-hour sessions are to be preferred over one-hour sessions for the purpose of the assessment are (1) data collection costs to the government are smaller per examinee-hour and (2) it would be difficult to cover the range of subject matter areas and also collect important background information in one hour. The incremental burden is from 25,000 student hours to 40,000 student hours, but this needs to be placed in the perspective of the 3 billion instructional hours that occur nationally at each grade level. The argument that the additional hour is a significant loss to the individual students is partially countered by the fact that the experience of the assessment may itself be educational. More important is the increased burden on schools' scheduling that may occur when planning for a two-hour session instead of a one-hour session. To minimize this burden, care must be taken to work with schools at an early point to set up convenient schedules. School administrators we have talked to agree that the difference between one- and two-hour testing sessions is an important but not crucial factor in deciding on participation in NAEP and that two hours would be quite acceptable if a reasonable rationale were presented.

One particular cost that has been suggested to be associated with testing across subject matter areas is the increased porportion of time needed for instructions. This position is questionable, however, because to a great extent instructions focus on the format of exercises, not their

content. It should be possible to define exercise formats so that the same formats occur across the various topics. Multiple-choice items in biology, health, history, and reading inference skills should not require separate instructions. Thus, even though several content areas are contained in an exercise booklet, as long as the exercises are grouped by format, little or no additional instructional time should be needed.

Finally, we need to consider costs associated with analysis and reporting. For the primary assessment report, the costs should be decreased, not increased, because the format and content will remain constant across assessment years. On the other hand, if one wishes to focus on a particular subject matter area, more powerful analyses can be done by combining data across assessments. This is essentially an extension of "matrix sampling" to the dimension of "years," and it requires the same type of combination algorithm as the other score generations based on matrix sampling. Therefore, we do not expect additional analysis costs to be large.

Conclusion

From this evaluation, we conclude that the benefits of this design change to NAEP significantly outweigh its costs. In addition, integrated annual assessments would be especially beneficial when combined with certain other design and procedural changes discussed in this report (for example, issue-based weighting of objectives, psychometrically sophisticated item development and analysis, or computer-assisted testing).

Measurement Founded on Modern Psychometric Theory

Background

When the National Assessment was being planned in the late 1960s, it was realized that traditional approaches to behavioral measurement would have to be modified very substantially to realize the goals of the assessment. Conventional sampling and psychometric techniques used commonly in pychological and educational measurement are designed to assess individuals on some psychological variable or achievement construct (see Lord & Novick, 1968). Examples of the former are measures of locus of control (e.g., Rotter, 1966) or test anxiety (Mandler & Sarason, 1960). Examples of the latter are typical vocabulary or arithmetic tests used to assess attainment of individual students.

The founders of NAEP were correct in realizing that the National Assessment should not rely on conventional psychometric methods. However, they were not successful in developing an alternate methodology that solves the nonstandard psychometric problems that an assessment presents. Furthermore, piecemeal modifications to the original strategy introduced in recent years are equally unsatisfactory. The following discussion describes the salient characteristics of the national assessment, the methodological problems these characteristics pose, and two attractive solutions to these problems.

The goals of a national assessment differ from conventional testing in at least two important ways. First, measurement at the level of the individual is not a goal; results are not to be used in making decisions on an individual examinee. Instead, an assessment should be designed to assess a group or an aggregate so that decisions about the progress of the group as a whole can be made. Ultimately, the "group" to be evaluated is the entire nation; however, smaller units of aggregation such as regions of the nation, states, districts, and types of schools are also of interest.

The second important difference concerns the type of material to be assessed. An assessment of educational progress should not be primarily concerned with measuring basic psychological constructs such as intelligence or spatial ability. Rather, an assessment is more properly concerned with measuring attainment in a large number of specific skill areas that make up the curriculum in the schools or are thought to be important for functioning adequately in society.

Aptitude versus Achievement

In the educational literature, this distinction is made between measures of aptitude and achievement (see DuBois, 1969; Green, 1974; Snow, 1980). Although the difference between the two concepts is by no means clear-cut, achievement, in general, refers to degree of mastery of some specified performance, while aptitude refers to an individual's ability to learn in the future. Some have thought of aptitude as stressing inherited ability, ease of acquisition, or relative fitness. Others have called aptitude "generalized achievement" and emphasized capacity to learn, solve problems, and reason logically.

It is clear that a major focus of any educational assessment should be on achievement rather than aptitude. Rather than reporting only on a single generalized achievement score, such as the verbal SAT, an assessment must report on attainment in each of a number of diverse skill areas. If there is any proper analogy to the Scholastic Aptitude Test, it is to their so-called "advanced" tests in specific content domains (e.g., foreign languages, physics, advanced mathematics) rather than to the two (much more publicized) basic aptitude scores.

The psychometric problem this creates is that the assessment must cover a highly multidimensional space and report on very specific content areas. Because an assessment does not focus on a few basic psychological variables, the conventional psychometric model of items as multiple indicators of a single ability may not be appropriate. In fact, items appropriate for an assessment can be highly curriculum dependent. That is, a

correct response to an item by a student may depend much more strongly on the instruction the student has had rather than his or her "ability." Thus, a psychometric model that presupposes that a correct response is a function of the examinee's "latent ability" may be less applicable to the psychometric problems posed by assessments than to the problems posed by the measurement of aptitude or ability.

Multiple Matrix Sampling

These differences led the founders of NAEP to construct a design that differed radically from the design of a conventional large-scale testing program. Perhaps the most dramatic difference is in the area of sampling. Because the assessment of individual students assumed no importance, it was decided to employ nonoverlapping multiple matrix sampling techniques rather than conventional examinee sampling. With this approach, each examinee responds to only a few items corresponding to a particular objective but responds to a broader range of items than would be possible with examinee sampling. The latter feature is important, given the requirement that the assessment report on attainment in many diverse content areas.

Used in the context of assessments, multiple matrix sampling has several important advantages. First, proper execution of item-examinee sampling will yield more precise estimates at the group level than will examinee sampling. Second, it facilitates the administration of a wider variety of items within fixed time constraints since each examinee does not have to respond to all items in the entire assessment. Third, it lessens response burden on the schools and students and serves to lessen fears among students that the results will be used to evaluate them individually.

The use of multiple matrix sampling is unfamiliar to most educators and researchers, however, and its use creates several added complexities. The most important is that the increased precision for measuring attainment at the group level can only be realized if the appropriate estimates

are computed from several matrix samples. This procedure is unfamiliar to most researchers, who are familiar only with constructing scales within a single matrix sample (i.e., dataset). The necessity of constructing a scale from items in several matrix samples is hard for most persons to grasp. To illustrate this point, consider the following example.

Suppose a researcher is interested in measuring some achievement variable (say addition of fractions) at the classroom level. Fifty classrooms constitute the primary sampling units. Within each classroom, ten students are randomly selected, and each student responds to one randomly assigned item out of a set of ten items. Each classroom responds to all items, but each student only responds to one (randomly assigned) item. In this instance, it is clear that this design assesses classroom achievement more precisely than a design in which one student per classroom responds to all items. However, the logic of the matrix sampling design requires that the researcher assemble each classroom's score from the responses of each of its ten students. The score for each classroom is composed of the ten item-examinee samples administered within that classroom. What should be reported as an estimate of the score for the population of classrooms is one score based on five hundred individual responses, each to one of the ten related items. The relevant components of this single estimate are the scores from each of the fifty classrooms that constituted the primary sampling units in the study.

Although this is not a difficult procedure, most educational researchers are unaccustomed to it. In fact, the equivalent of this has not been done by the National Assessment. Instead, responses at the level of the item within a matrix sample, are reported. In our example, what would have been reported by NAEP would be the proportion of students responding correctly to each of the ten items. This would produce ten statistics, each based on fifty scores (one student per classroom).

It is apparent that this procedure does not capitalize on (1) the fact that a ten-item scale measuring one variable exists, (2) precision at the classroom level has been maximized (for the ten-item test), or (3) precision for the overall estimate has been maximized. In fact, the current

NAEP procedure of within-booklet (i.e., within matrix sample) reporting undermines a major technical advantage of matrix sampling and leads to fragmentation of the results of the assessment. The latter point is discussed in greater detail below.

Reporting Results from Individual Exercises

In response to the need of the assessment to focus on achievement rather than aptitude, the original planners of NAEP decided that the reports on educational progress should take an unusual form. Rather than constructing any type of scale score, NAEP reports would be written in terms of responses to each individual exercise. Such a strategy is consistent with the goal of reporting in very specific content areas but has some severe limitations. First and foremost, for the meaning of the percentage correct to be interpretable, the exercise must have an importance that is self-evident and unambiguous to the reader. Such conditions are approximately met in opinion research, such as the Gallup poll and its competitors. For example, questions pertaining to voter preference in a specific election have meaning on their own, without appeal to any psychological construct (e.g., liberalism or conservatism) or as being representative of some domain. This so-called "fixed-item" approach works well in social survey research in which the responses to items may be interpreted at face value. This is especially true when the question pertains to some particular action the respondent may take (e.g., voting preferences, response to a draft notice). In such instances, the question is not thought of as one of a larger universe of questions.

Very few educational test items have such singular importance, however. More typically, test items are interpreted as representatives of a population of test items that could be written to assess a particular skill. The crucial point is that no single item is accepted as the definition of the skill. Instead, it is accepted that several items define a domain and that examinees respond probabilistically to these items (i.e., some examinees may get an item right due to guessing, and others may get an item wrong due to carelessness). Thus, a strategy of reporting at the

item level is fraught with interpretational difficulties. In the area of the NAEP mathematics assessment, Haertel (1981) posed the basic issue:

> Only 7% of seventeen-year-olds could correctly solve the equation $(x-2)^2=9$ for x, but in another sample, 18% could "find the solution set of $x^2-5x+6=0$." Is the difference due to the wording of the problems? The particular numbers? The format of the equations? How are we to generalize about the proportion of seventeen-year-olds who can solve quadratic equations? How would the p-values change if these items were multiple choice, say, rather than free response? There is no way to tell.

On the one hand, it is clear that the percentage of correct responses reported for exercises such as these conveys little, if any, meaning since those percentages are a function of the format of the exercise, distractors used, and difficulty of the particular exercise stem. But on the other hand, the public and the educational community is interested in knowing about level of achievement in very specific content areas such as "solving quadratic equations." The requirement that an assessment report on attainment in many skill areas does not, in fact, free it from the requirement that those reports be in some interpretable metric that is invariant with respect to choice of exercise set within a skill area.

Latent Trait Analysis

An important challenge to NAEP is to develop and use a methodology capable of reporting in specific skill areas without becoming tied down to specific exercises. In fact, what is needed is a methodology that directly and unambiguously addresses questions such as "How are we to generalize about the proportion of seventeen-year-olds who can solve quadratic equations?". Such a methodology should be capable of addressing questions phrased in terms of the skill areas themselves, independent of particular exercises chosen within a domain, and produce estimates invariant with respect to the exercise-examinee sampling procedures.

A natural place to begin development of such a methodology is latent trait or so-called "item response" theory. Originally developed in the context of the measurement of individuals, this family of models can be used to produce scale scores in an arbitrary metric in interval scale units. Unidimensional exercise sets are produced in the test development stage and calibrated in preliminary studies. Once the characteristics of the exercises are known, any subset of exercises in the item bank can be administered and individuals' scores on the latent trait can be estimated from the results. These scale scores are comparable even if some examinees get different subsets of the exercise set than others.

Borrowing from item response theory, Bock (1976, 1981, 1982) and his colleagues (Mislevy, Reiser, & Zimowski, 1981; Reiser, 1980) have made considerable progress in adapting latent trait methods to the unique problems posed by assessments. As is implemented in the current design for the California assessment, scale scores in each of more than sixty specific skill areas are computed using latent trait models modified to handle multiple matrix sampling. In this way, scale scores in the domains of interest can be reported without depending on specific items. The methodology depends only on the development and maintenance of a bank of calibrated exercises in each content domain. New exercises can be added to the assessment as old ones are released to the public without compromising in any way the ability of the National Assessment to measure change. Furthermore, results of the assessments can be reported directly in terms of the skill areas of interest rather than in terms of specific exercises, whose coverage of the skill area is incomplete and whose psychometric characteristics are unknown.

The conventional machinery for latent trait estimation must be generalized to handle the complexities created by multiple matrix sampling. Since exercise-examinee sampling procedures may dictate that any one student only takes a very small number of exercises per skill area, latent trait estimation at the level of the individual is very imprecise. What is needed is a methodology for defining the latent trait at the group level instead. Such a generalization of Bock's (1976) model was performed by Reiser (1980). In Reiser's model, the probability of a correct re-

sponse to a particular exercise by a student selected at random from the group is a function of the exercise parameters and the average level of attainment in that group. The latter is a function of the main effects and interactions that define that group. His estimation procedure produces the information about the population and subpopulations that the assessment is designed to provide. An advantage of Reiser's procedure is that it is designed to produce scale scores for a population from as few as one exercise per skill area per booklet.

In the California assessment, each of 62 skill areas is assessed using an average of sixteen items each. However, each examinee responds to no more than two items per skill area. Using a latent trait model generalized for group data, scale scores at the unit of the school are reported. As is the case for latent trait methods designed to score individuals, the method produces scale scores in well-defined units suitable for the measurement of change. Mislevy, Reiser, and Zimowski (1981) used this procedure to study change in mathematics attainment from 1972 to 1977.

Using such procedures, scale scores can conveniently be computed from exercises in several matrix samples. Because the exercises have been calibrated using the latent trait methodology, these scales are invariant with respect to addition or deletion of particular exercises defining that skill area.

Latent Class Analysis

Although this generalization of item response theory is a very marked improvement over the present practices of NAEP in that it uses information from all booklets in an efficient manner and reports in terms of scale scores, its applicability to highly curriculum-dependent types of exercises is open to question. For such exercises, the dimensionality of the space is obviously greatly affected by whether or not students have received instruction in the area the exercise assesses. Thus, both the patterns of interitem association and difficulty level are strongly

affected by the school curricula, rather than merely by the ability (aptitude) of the students. Indiscriminant use of methods designed primarily to assess stable characteristics of the person can be misleading if used in this context.

As an alternative to use of latent trait methods in an assessment context, Haertel (1980, 1981) has proposed that restricted latent class methods (Lazarsfeld & Henry, 1968) be used to model the item response data. The advantage of these methods is that their assumptions are likely to be more congruent with the nature of highly curriculum-dependent item responses. A distinctive feature of these models is that skills are treated as dichotomous: A given examinee either does or does not possess each skill. If an item requires only the skills an examinee possesses, then he or she can solve the problem; otherwise, he or she cannot. If such models are applied to data arising from studies designed to assess individual performance, this might not be an appropriate assumption. But for describing populations, the models work well. The assumption of skill dichotomies corresponds naturally to the fact that responses are strongly influenced by whether or not students have received instruction in the skill area. The methodology of the latent class analysis itself is required to cope with the probabilistic nature of item responses.

The methodology developed by Haertel for analyzing assessment data involve the following steps:

First, exercises are characterized according to the skills required to successfully solve them. Unlike latent trait analysis, which ordinarily requires that the responses are a function of only one latent trait, latent class analysis permits the researcher to study exercises that require several skills for correct response.

Second, the union of all skills needed to solve all exercises is assembled. Each subject, then, is assigned to some skill profile based on that subject's pattern of right and wrong responses to the set of exercises.

Third, the statistical analysis assigns some probability to each possible skill profile. The probability is interpreted as the estimate of the population proportion that possesses that pattern of skills. The analysis is probabilistic: it recognizes that the skills an exercise requires and an examinee possesses are the sole determinants of the probability that the examinee will answer correctly. In fact, there is some (hopefully low) probability that an examinee lacking one or more of the requisite skills will answer the item correctly and the (hopefully high) probability of a correct response by an examinee who possesses all the requisite skills. These are known as the false positive and true positive rates, respectively. The former probability is, in general, greater than zero due to guessing, and the latter is, in general, less than one due to carelessness. The analysis consists of estimating simultaneously the proportions of examinees in each latent class (i.e. skill profile) and the false positive and true positive rates for each exercise part.

The great advantage of this method is that these proportions are descriptive of component skills rather than specific exercises. Furthermore, they appear to possess the desired properties of invariance across examinee-exercise sampling that is crucial if such statistics are to be meaningful. In his study of math attainment, Haertel (1981) found that his estimates were not significantly different across NAEP booklets, and thus, could be combined to produce estimates for the population as a whole. Due to the structure of NAEP matrix sampling, this invariance held across examinee-exercise pairings; this is obviously the most stringent practical test of invariance.

Like the latent trait methodology, the latent class approach frees the National Assessment from reporting merely single exercises, but in addition it permits a more fine-grained and theoretically defensible analysis of attainment in many different skill areas. Most notably, latent class analysis is very well suited for the analysis of patterns of skill acquisition, since it explictly takes into account the fact that several skills may be required to respond correctly to a given item. Ordered or hierarchical patterns of skill acquisition may conveniently be studied.

The results are especially amenable to description in simple declarative sentences because the report is worded in terms of the proportion of the population possessing a given skill or pattern of skills (e.g., "Thirty-five percent of seventeen-year-olds could solve linear equations in one unknown"). That is, the skill dichotomy assumption corresponds well with the layman's notion of skill mastery. The lay public can easily understand the meaning of statements like "XX% of seventeen-year-olds can balance their checkbooks" or "...can understand labels on products in the grocery store." Such statements actually invoke the concept of the latent class and the idea of generalization across both stimuli and time.

It is curious that the California assessment attempts to meet this need within the context of latent trait methods. To define "mastery," Bock (1981) arbitrarily chose an 80% probability that a randomly selected student would get an item right. He then can report on the proportion of students who have reached "mastery." Of course, another arbitrary choice of mastery level would produce different estimates. Clearly, it is preferable to use an analysis that defines mastery level empirically. The methodology developed by Haertel accomplishes this.

Conclusion

Although the final results of a latent class analysis are worded in simple language, the technical problems involved in generalizing from specific exercises to component skills are far from trivial. Regrettably, the current simple-minded policy of item-by-item reporting to address specific skill areas is not acceptable; to meet the original goal of reporting on progress in many diverse skill areas, it is essential that the modern psychometric techniques discussed here be brought to bear on the problem.

In our view, latent trait analysis has a more limited place in educational assessments than does latent class analysis. Latent trait analysis lends itself much more readily to the measurement of higher level and more generalized cognitive skills than of the present objectives of NAEP.

However, it is clear that such variables do have a place in a National Assessment. The prominence of Scholastic Aptitude Test scores suggests that such generalized variables, in fact, have considerable impact with the general public. Similarly, the research commmunity has shown that aptitude variables are highly relevant to policy questions (e.g., Cronbach & Snow, 1977). There is no doubt that the generalizations of latent trait techniques for group data due to Bock and his colleagues are the methods of choice for aptitude indicators. For analysis of national assessment achievement data, however, the latent class methods provide a far more technically defensible and readily interpretable means of reporting in specific skill areas.

Computer-Administered Testing

Feasibility

The basic questions of whether computers can be used to administer tests and whether adaptive testing is technically feasible have already been answered positively. For a number of years, Frederic Lord and David Weiss, among others, have been doing research on these topics, and efficient procedures have been developed for using computers, whether they be mainframe, mini, or microcomputers, to administer tests. The flexibility of test presentation, control of administration, and sensitivity to responses promised by a future of computer-administered tests calls for the kind of leadership in technical innovation for which NAEP was designed. It is altogether appropriate, we believe, for NAEP to aim towards a goal of computer-based assessment.

Practicality

A question that has not been answered, however, is whether computer testing is yet a practical approach to testing, especially for an effort the size and scope of NAEP. Virtually all studies of computer-administered testing, especially adaptive testing, have focused on tests composed of multiple-choice items where each item is independent of all other items; and where the question, all options, and any associated stimulus material can all be displayed together on a single CRT screen display. The examinee makes his or her choice based on the material shown on the screen and then the computer selects the next item to be presented, either adaptively or in sequence, and presents that material. Multiple questions based on the same stimulus materials have been little used in studies of computer-administered testing, nor have items that are based on figural or long textual materials. Item formats other than multiple choice have received little or no attention. While these limitations would place only small restrictions on NAEP exercises in certain content areas, it is clear that many NAEP exercises, perhaps even whole content areas, could not be administered under these restrictions.

In addition to limitations in the types of test items that have been used in computer-administered testing, computer testing has still not been implemented widely, particularly at the elementary and secondary school levels. Even the military services, usually leaders in the adoption of new technology, have not yet made any large scale commitments to computer testing. There is currently a program underway to computerize the administration of the Armed Services Vocational Aptitude Battery (ASVAB), which is administered to all potential enlistees. This effort is still at least several years from implementation, however, and is intended to be limited to multiple-choice items. While the PLATO system might appear to be a major user of computer testing, the testing is in fact an integral part of the individualized teaching/learning system and not a stand-alone testing program. In terms of stand-alone computer testing, the largest use to date may in fact be with specialized microcomputer systems designed for use in the offices of counselors and clinical psychologists. These computers administer, score, and provide rapid feedback of results on tests and inventories such as the MMPI.

There appear to be two major reasons for this lack of widespread adoption of computer testing.

- To date, the hardware costs associated with implementing such an approach to testing have been quite high, although costs continue to decrease rapidly as the technology evolves.
- Computer output format restrictions in terms of what can be displayed on a CRT screen have, for the most part, limited the content to stand-alone, self-contained multiple-choice items.

If these two barriers can be removed, and we feel that there is a high potential for this happening, then there is a bright future for computer testing in many areas of education, including NAEP.

Recommended Feasibility Study

We recommend that a two-pronged effort be undertaken to determine the potential for application of computer testing to NAEP and to monitor changes in that potential over time. One aspect of this effort would involve negotiating with computer hardware developers to identify advances with the greatest potential either to decrease the cost or to increase the performance of computer testing systems. To accomplish this end, the NAEP grantee should establish a semi-formal communications network with hardware experts from various segments of the computer industry. Working with the advice and counsel of these individuals, particular attention should be paid to such issues as:

- Should computer testing systems be developed around mainframe computers, minicomputers, or microcomputers?
- Should a testing station consist of a terminal linked to a master computer, or should each station be a stand-alone computer?
- If stand-alone computers are used, should each station have independent memory storage, or should they be linked to a common storage device such as a Winchester Disk?
- What types of display options could be used? What about the use of videodiscs for image storage and display?
- How does the current cost of an optimal system compare with the printing and scoring costs that would be saved?

The other part of this approach should involve a small-scale (relative to the total NAEP budget) research effort to seek answers to some of the practical questions related to the use of computer testing in NAEP. The following issues should be investigated.

- What is the most effective way to administer various kinds of exercises that are not multiple choice in format? This would include exercises that require that the examinee fill in missing information, write sentences or paragraphs, produce or perform some work, and so on. Would tests of response speed add important mediator information not available from paper-and-pencil tests?

- What is the best way to present long or complex stimulus materials such as reading passages, charts, figures, and the like? Finding a satisfactory answer to this question is critical to the presentation of many NAEP exercises via computer. For example, as a part of an art exercise, the examinee might be required to look at a detailed or colored drawing or picture, which would require the use of a color display device. Similarly, as a part of a music exercise, the examinee might be required to look at part of a musical score or listen to a passage, which would also require a special display device.

  The same sort of display factors apply to reading exercises where the examinee is to read a passage and then answer questions about it. In a typical paper-and-pencil reading test, the entire passage and all of the questions are available to the examinee at all times, and the examinee is free to look back at the passage as often as is desired. In fact, the most effective strategy for answering questions about a reading passage is to read the questions before ever looking at the passage and then read the passage to find the answers to the questions. With a computer-administered test, however, it is likely that more than one complete CRT display would be required to present the reading passage and an additional CRT screen display would be required for each question. While the examinee could jump forward or backward among the text passage and question screens by pressing keys, this is not the same as scanning back and forth by eye. NAEP needs to determine whether computer-presented exercises are equivalent to the same exercises presented in exercise booklets in terms of examinee scores, reliability, and interactions with examinee characteristics. If there is a non-equivalence, it is possible that the computer-based items will be found to be more powerful. However, when comparisons are to be made with previously administered paper-and-pencil tests, appropriate adjustments may be necessary.

- What about the use of computer-controlled videotapes, videodiscs, or slide projectors to display needed information on a small screen or a second CRT display in place of a supplementary printed information booklet? While technically feasible, at the present time these approaches are rather expensive--so much so that their use would probably not be cost effective now. However, NAEP staff should continue to monitor developments in display technology so as to be aware when there are significant cost decreases in current technology or when new technology develops. For example, although laser videodisc mastering costs $2000 at present, it will cost $20 with technology currently on the drawing boards.

- Are scores obtained by computer testing and paper-and-pencil testing approaches equivalent? For example, we know that some examinees experience test anxiety and that there are some individuals who experience computer anxiety. Will the combination of computers and tests result in greater levels of anxiety and thus lower test scores? Will the introduction of computer testing cause Hawthorne effects to occur (i.e., where the novelty and perceived special attention causes the examinees to try harder and perform better)?

- How will computer testing influence the test-taking strategies of examinees? For example, when answering a paper-and-pencil test, many examinees skip the items they find to be more difficult and come back to them later. What is the appropriate level of control of this behavior to impose with computer administration?

- What are the logistics of temporarily placing computers at schools for testing? Can this be done by local school personnel, or will NAEP-trained personnel have to continue going from school to school to conduct the testing? (If the computer system required little supervision, the hardware and developmental costs would be partially balanced by decreased costs for administration personnel, especially for individually administered exercises.)

- What are the logistics of combining data obtained from the many testing sites so as to produce the final data files upon which results and reports will be based? How much aggregation will take place in the field, possibly during the testing, and how much will be performed at a central location?

- Should NAEP consider using branching testing in which answers to certain background questions will determine which of several tests an examinee should take? In foreign languages, for example, it would make sense to give a French test only to individuals who had studied French and a Spanish test only to individuals who had studied Spanish. The same approach might apply with a mathematics test, especially at the senior high school level. The examinee who had studied advanced algebra, solid geometry, or even the calculus might receive a different test from someone who had only taken general or business mathematics. Use of computer testing makes branching like this highly feasible.

To obtain answers to these questions, and others that are likely to arise, the NAEP grantee should acquire a prototype multi-examinee computer testing system, which can be used to carry out empirical studies. Prior to acquiring such a computer testing system, however, NAEP staff should

consult with experts in the computer hardware field to seek their guidance with regard to anticipated future developements. The system acquired should be one that best anticipates and can take advantage of future developments in terms of cost and capabilities. It is possible that a customized, rather than an off-the-shelf, system would be the most cost-effective approach to computer testing in NAEP. Such a customized system would contain only those features required to carry out the testing and would omit other costly features that are not required.

The first studies would explore general questions like those mentioned above. Then, in the 1985-86 assessment, a particular small subset of NAEP might experimentally be performed using computers. The outcome of this demonstration would guide subsequent expansion or revision of plans for computer administration.

Conclusion

We have no doubt that eventually the National Assessment will be administered largely or entirely by computer, with great increase in flexibility, efficiency, and the amount of information collected per examinee-hour. The primary questions are merely when and how. We believe that the gradual exploration and introduction of computers into test administration is the optimal form of leadership role for NAEP to play in this area.

Establishment of an Educational Assessment Institute

Introduction

The NAEP grant from NIE will only provide limited resources for scholarly inquiry into large-scale assessment methodology and utilization, and it will provide no resources for independent monitoring and critiquing of NAEP policies and procedures. As a result, AIR's proposal for a NAEP planning grant included brief reference to an independently-funded Educational Assessment Institute that would support joint reseach and development activities aimed at improving large-scale assessment theory and practice. Our planning grant activities focused initially on determining the functions such an institute might carry out and how it might be organized. When collaborative activities with the Stanford University School of Education produced an apparently useful blueprint, we followed up with preliminary explorations to identify (1) requirements and sources for independent funding and (2) initial planning steps that would be required to tap these sources and establish the Institute. This chapter provides an overview of the Institute's potential functions, organization, and funding. It concludes with a report on the current status of Institute planning efforts. The orientation of this chapter is necessarily centered on the Stanford/AIR locus of the Institute, but the concepts might equally well be applied to another grantee. We expect that the Stanford/AIR locus will prove attractive to potential funding sources.

Potential Functions for an Educational Assessment Institute

AIR and the Stanford University School of Education jointly identified five major functions that an independent Educational Assesment Institute might carry out: (1) support resident scholars, (2) independently review and critique NAEP policies and procedures, (3) conduct research and training seminars, (4) sponsor an annual conference on large-scale assessment, and (5) interact with the international assessment community.

Resident center for scholarly inquiry. Based on the model provided by the Center for Advanced Studies in the Behavioral Sciences, which is located on the Stanford University campus, the Educational Assessment Institute would be a resident center for scholarly inquiry. It would provide six-month and twelve-month resident fellowships for researchers who merit these appointments. Fellows would be awarded both living expenses and stipends in lieu of their regular salaries. Nominations for fellowship recipients would be solicited from the general academic community; nominees who pass initial screening criteria would be asked to submit applications describing their research interests in the Institute's two major fields of inquiry--large-scale assessment methodology and techniques to promote utilization of assessment information for improving education. Applicant credentials and statements of research interest would be reviewed semi-annually by an independent Institute Board composed of eminent scholars drawn from the Fellows of the National Council for Measurement in Education (NCME) and Divisions 5 (Measurement) and 15 (Educational Psychology) of the American Psychological Association and from the senior AIR staff and faculty of the Stanford School of Education. Selected applicants would normally take sabbatical leave from their current positions to spend a specified period (six or twelve months) in Palo Alto under the auspices of the Educational Assessment Institute. While in residence, fellows would have access to Institute-funded computer and clerical support and all NAEP public-use resources, including the NAEP item and data banks, Clearinghouse, and computer software packages. They would also interact with senior AIR and Stanford School of Education staff through frequent in-house seminars.

Permanent center for scholarly review and critique of NAEP policies and procedures. In their recent volume Toward Reform of Program Evaluation, Cronbach and his associates (1980) presented two general theses that are relevant to this proposed function:

> Oversight by peers is the most promising means of upholding professional standards and of precipitating debate about strategic and tactical issues.

And:

> The best safeguard against prematurely frozen standards for evaluative practice is multiple, independent sources of criticism. (p. 10)

Several recent examples illustrate the practical value of these theses in relation to major educational program evaluations. In the early 1970s, the Huron Institute was requested to provide planning and monitoring assistance to U.S. Office of Education staff who were initiating the major Follow-Through planned variations study. Huron's role evolved over the years into that of scholarly critic; moreover, this role was influential in helping to shape the Follow-Through evaluation in numerous ways. As Michael Garet's final report evaluating Huron's work makes clear, "measured against technical, organizational, political, and social-scientific criteria, Huron's impact on the evaluation in the final years was without doubt a healthy one" (Garet, 1978, p. 68). Huron became a kind of broker as well as being a source of bright ideas, technical advice, and criticism--it smoothed communication between various parties involved in the evaluation, exerting quiet influence both in Washington and in the evaluation contractors' offices. Garet (who is now on the Stanford University faculty) made important recommendations about how, in the future, monitors/critics might be selected to work on major evaluations; the language needs only minor editing to apply to an independent educational assessment institute charged with infusing NAEP with fresh ideas, perspectives, and constructive criticism:

> There are several criteria that might be considered in selecting an organization to serve as monitor....
>
> 1. The technical skill of the external monitor should, of course, be the equal of that of the evaluation contractor.
>
> 2. The monitoring organization should possess the ability to work closely with the major organizational units involved in conducting the evaluation as well as the flexibility to shift resources easily from one monitoring task to another....
>
> 3. The organization should have a certain amount of legitimacy among the evaluation constituencies; that is, it should

hold a secure, independent status based upon a serious and continuing interest in the problems and programs being evaluated....

4. The organization should have a relatively strong research identity; that is, it should have a fairly coherent social-scientific approach to the problems of evaluation, in order to encourage a meaningful dialogue concerning evaluation methods and results. (pp. 72-73)

More recently, Charles Murray (1980) made similar observations about the role played by an Evaluation Research Society (ERS) panel in shaping AIR's NIE-sponsored evaluation of the Cities In Schools service integration experiment. The independent ERS panel, constituted prior to the evaluation and maintaining ongoing contact with it, was thought to be much more effective in giving useful (and heeded) advice than would have been the case with one-shot, usually post hoc reviews by more traditional scholarly critics.

We anticipate that this function could be one of the most important to be carried out by the NAEP Educational Assessment Institute and would be the one to which permanent Institute-affiliated staff from both Stanford University and AIR would devote a high proportion of their efforts.

Center for research and training seminars. The Educational Assessment Institute would periodically organize and sponsor research and training seminars (usually in cooperation with other scholarly organizations and institutions of higher education). The purpose of these seminars would be to share information about the techniques and findings of ongoing research into large-scale assessment topics. Institute staff and fellows would organize the seminars. Participants would be assessment researchers around the country who could benefit both by learning new techniques and by receiving peer criticism of their own work.

As a general rule, seminar participants would be required to pay their own expenses. If desirable, continuing education credits could be awarded by the Stanford School of Education. Seminars would be announced through mailings to members of national measurement and evaluation profes-

sional associations. (Especially well-received seminars might also be replicated at professional association conventions.)

In cooperation with other NAEP efforts to utilize new communication technologies, the Institute would also sponsor an ongoing large-scale assessment seminar via teleconferencing. Participants would all pay a small fee and would then be given access to a computer-based telecommunications network through which documents, questions, probes, thoughts, dialogues, and group conversations could be shared, recorded, processed, and analyzed. All researchers with access to a microcomputer or an appropriate computer terminal with a telephone modem could participate in this network without leaving their offices. Recent experience with such ongoing teleconferences has begun to demonstrate their potential as useful and inexpensive tools for timely problem identification and definition, solution building, and policy evaluation. For example, the SPECIALNET teleconference network now operating under the sponsorship of the National Association of State Directors of Special Education (NASDSE), serves the following functions:

- providing electronic mail service, including person-to-person messages and group announcements;
- obtaining immediate feedback from all or predetermined representative samples of state directors regarding questions of interest to network members, including those related to possible national policies;
- sustaining ongoing seminars of interest to specialized subgroups of network members;
- facilitating short-term collection of evaluative data and feedback of results for individual network members or the NASDSE; and
- providing easy access to computer utility functions such as report generating and word processing programs for all network members.

AIR is operating a similar network, the VIM Network, to facilitate coordination and evaluation of interactive videodisc use in basic skills instruc-

tion, under the sponsorship of the Division of Educational Technology, U.S. Department of Education.

Sponsor for the Annual Conference on Large-Scale Assessment. Acting together with the NAEP project staff, Stanford Univerity, and measurement and evaluation professional associations, the Educational Assessment Institute would assume sponsorship of the Annual Conference on Large-Scale Assessment. As it does now, the conference would focus on major research issues regarding assessment techniques and information utilization. When they register for the conference, participants would be asked to nominate research topics they hope to see on the conference agenda. A team of national and state assessment experts would then be invited (and paid) to attend the conference and lead discussions on those topics and others determined by the Institute Board.

The 1982 Twelfth Annual Conference in Boulder, Colorado, attracted over 260 attendees who were representatives of state and local education agency evaluation offices, college and university faculty, and national professional associations. Future annual conferences, which we recommend be held in Washington, D.C., would be expected to attract an equally large or larger number of participants, especially in view of the fact that attendees would be able to help structure the conference agenda.

U.S. locus for liaison with the international assessment network. A small but active network of international assessment scholars has sprung up over the last 15 years. This network is headquartered at the International Association for the Evaluation of Educational Achievement in Stockholm. In the past, U.S. participation in the Association has largely been on an ad hoc basis, with no permanent locus for organizing U.S. participation.

The Educational Assessment Institute would seek to organize all U.S. participation in the International Association by providing staff time, access to the NAEP exercises and data base, and logistical support for U.S. contributions to international assessment endeavors. If sufficient

funding can be located, Institute scholars might participate in or even organize international data exchange, analysis, and reporting efforts.

Institute Organization

The proposed Educational Assessment Institute would eventually be incorporated as an independent not-for-profit organization under the leadership of a permanent staff and Board. As mentioned previously, this Board would be composed of nationally-recognized assessment scholars representing major professional associations, Stanford University, and AIR. During its initial start-up phase, the Institute would probably be organized as a center under the aegis of the Stanford School of Education, and its initial Director would be selected from among the faculty (active or emeriti) of that school. After the Board had organized, it would select a permanent Director, who, in turn, would select the Institute staff. We anticipate that some staff affiliations might be part-time, allowing access to Stanford University faculty and AIR professional staff having other research and teaching commitments.

The Institute would be physically situated either on the Stanford University campus or in quarters near the campus, providing easy access to both the School of Education and AIR.

Institute Funding

We assume that initial support for Institute planning and fundraising activities would be provided by NIE through a modest line item in the NAEP contract budget. These initial NAEP planning funds would be used to seek permanent funds from private foundations having interests in the improvement of American education.

The initial planning period would extend approximately eight months, during which the following activities would be carried out, under the overall direction of the Dean of the Stanford University School of Education:

- appoint an interim Director;
- solicit nominations for and select the Institute Board;
- prepare an Institute prospectus and circulate it widely to private funding sources with which contacts have already been established through previous support;
- prepare a detailed proposal for $500,000 to provide three years of core support and submit it to those foundations expressing interest in the preliminary prospectus; and
- obtain funding to establish the Institute, with initial functions and staff size to be determined by the level of funding achieved.

Present Status of Institute Planning

The Stanford School of Education has already drafted and prepared for circulation a preliminary Institute prospectus. This prospectus will shortly be sent to senior staff of several foundations known to have priorities in related areas. In the event expressions of interest are received as a result of these preliminary inquiries, NIE will be immediately notified.

References

Bock, R. D. Basic issues in the measurement of change. In de Gruijter, D. N. M, & van der Kamp, L. J. T. (Eds.), Advances in psychological and educational measurement. New York: Wiley, 1976, 75-110.

Bock, R. D., & Mislevy, R. J. An item response curve model for matrix-sampling data: The California grade three assessment. New Directions for Testing and Measurement, 1981, 10, 65-90.

Bock, R. D., Mislevy, R. J., & Woodson, C. The next stage in educational assessment. Educational Researcher, 1982, 11(3), 4-11.

Cronbach, L. J., & Snow, R. Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington, 1977.

Cronbach, L. J., et al. Toward reform of program evaluation. San Francisco: Jossey-Bass, 1980.

DuBois, P. H. (Ed.) Invitational conference on testing problems. Princeton, NJ: Educational Testing Service, 1969.

Education Commission of the States. A guide to National Assessment objectives and items (SY-OI-36). Denver, CO: Author, 1981.

Flanagan, J. C. The critical-requirements approach to educational objectives. School and Society, 1950, 71, 321-324.

Flanagan, J. C. (Ed.). Perspectives on improving education: Project TALENT's young adults look back. New York: Praeger, 1978.

Flanagan, J. C., & Russ-Eft, D. F. An empirical study to aid in formulating educational goals. Palo Alto, CA: American Institutes for Research, 1975.

Garet, M. S. External monitoring and the management of educational evaluations: Lessons learned from the Huron Institute's involvement in Follow Through. Washington, D.C.: U.S. Office of Education, 1978. (Contract Report OEC-0-74-0394)

Green, D. R. (Ed.) The aptitude-achievement distinction. Monterey, CA: CTB/McGraw Hill, 1974.

Greenbaum, W., Garet, M., & Solomon, E. Measuring educational progress. New York: McGraw-Hill, 1977.

Haertel, E. H. Determining what is measured by multiple choice tests of reading comprehension. Unpublisheed doctoral dissertation, University of Chicago, 1980.

Haertel, E. H. Developing a discrete ability profile model for mathematics attainment. Final report submitted pursuant to Grant NIE-G-80-0003 submitted to National Institute of Education, 1981.

Lazarsfeld, P. F., & Henry, N. W. Latent structure analysis. New York: Houghton Mifflin, 1968.

Lord, F., & Novick, M. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Mandler, G., & Sarason, S. A study of anxiety and learning. Journal of Abnormal and Social Psychology, 1952, 47, 166-173.

Mislevy, R. J., Reiser, M., & Zimowski, M. Scale score reporting of national assessment data. Final report to the Education Commission of the States under contract #02-81-20314. Chicago: International Educational Services, 1981.

Murray, C. A. The ERS review panel for AIR's Cities In Schools evaluation. Paper presented at the annual meeting of the Evaluation Research Society, Rosslyn, Virginia, 1980.

PLAN master objectives. Palo Alto, CA: Westinghouse Learning Corporation, 1971.

Reiser, M. A latent trait model for group effects. Unpublished doctoral dissertation, University of Chicago, 1980.

Rotter, J. Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 1966, 80 (1, Whole number 609).

Snow, R. Aptitude and achievement. New Directions for Testing and Measurement, 1980, 5, 39-59.

University of Texas. Adult functional competency: A summary. Austin, TX: University of Texas, 1975.

APPENDIX A

PLAN Master Objectives

January, 1971

77

Westinghouse Learning Corporation
PLAN Division

FOREWORD

A huge advantage of an instructional objective derives from the simple fact that it is written down. Once it is written, it is visible. Once it is visible, it can be reviewed, evaluated, modified, improved.

Objectives are frequently discussed, but seldom seen. In these volumes you can see some four thousand instructional objectives in the subject areas of Language Arts, Mathematics, Science, and Social Studies, extending over the range from grade one through grade twelve. This collection represents the cooperative efforts of over one hundred classroom teachers and an almost equal number of staff members at the American Institutes for Research and the Westinghouse Learning Corporation.

Since these volumes present written objectives rather than offer a discussion about objectives, they become the criteria by which materials are selected, content outlined, instructional procedures and educational technology developed, and tests and examinations prepared. All aspects of an educational program are really the means to accomplishing the basic educational purpose. This collection serves to stimulate teachers and educators in selecting and developing behavioral objectives for their local use. These objectives may be criticized and evaluated, revised and modified, additions made to or objectives deleted; all with the view of arriving at an appropriate set of educational outcomes to meet the educational needs of a local situation and of individual students.

The rather obvious purpose of an instructional objective should be to make clear to teachers, students, and other interested persons what youngsters should be able to do as a result of the instructional program. A well-written instructional objective should specify under what conditions and to what extent a certain kind of student performance can be expected to take place.

Unfortunately, school systems commonly lack a comprehensive and reasonably consistent set of educational objectives. Educational goals and objectives are quite frequently expressed only in broad, global terms, and the question of what and how to teach is left to a considerable extent to the teacher. As a result, quality in the schools is closely associated with the qualified and artful teachers. No doubt considerable excellent educational work is done by artistic teachers who, while they do not have a clear conception of goals, do have an intuitive sense of what is good teaching. Their materials are significant, and they develop topics effectively with students. The artistic teacher clarifies the educational objectives (even those not directly stated) through her actions as she teaches intuitively.

If the foregoing were to serve as a basis for defining education, then the "intuitiveness of the artistic teacher" would have to be built into the educational program. This, of course, cannot be done. The alternative is to start with clearly defined, rather than implied, instructional objectives.

Educational objectives—even clearly stated, specific objectives—are in the final analysis matters of choice and thus are value judgments. The question then arises:

> Who provides these value judgments? In the last analysis, the public schools are operated to meet the needs of society. Some of the objectives and who shall attend school are provided for in the state constitutions and by laws. Others are set forth by the efforts of the elected representatives of the people of a community. Others are provided by the professional educators hired to operate the schools.

Others come from our knowledge of the children themselves and how they learn. These effectively furnish the sources of educational objectives for a local public school. They will change with the changing conditions of the times; sometimes fast, as with Sputnik, but usually slowly.

In evaluating and summarizing instructional objectives, whatever their source, certain kinds of information and knowledge provide a more intelligent basis for making decisions about objectives. If these facts are known and understood, the probability is increased that judgments about objectives will be wise and that the school goals will have greater significance, objectivity, and validity. For this reason, a large part of the so-called scientific study of the curriculum has concerned itself with investigations that may provide a more adequate basis for selecting instructional objectives wisely.

The question is then raised as to what sources can be used for getting information that will be helpful. A good deal of controversy goes on between essentialists and progressives, between subject specialists and child psychologists, between sociologists and the philosophers, between this school group and that school group, over the question of the basic source from which objectives can be derived. The progressives and child psychologists emphasize the importance of studying the child to find out what kinds of interests he has, what problems he encounters, what purposes he has in mind. They see this information as providing the basic source for selecting objectives. The essentialists and subject specialists, on the other hand, are impressed by the large body of knowledge collected over many thousands of years, the so-called cultural heritage, and emphasize this as the primary source for deriving objectives. They view objectives as essentially the basic learnings selected from the vast cultural heritage of the past.

Many sociologists and others concerned with the pressing problems of contemporary society see in an analysis of today's world the basic information from which objectives can be derived. They view the school as the agency for helping young people to deal effectively with the critical problems of modern life. If they can determine what the existent problems are, then the objectives of the school are to provide these knowledges, skills, attitudes, that will help people to deal intelligently and effectively with contemporary problems. On the other hand, the educational philosophers recognize that there are basic values in life, largely transmitted from one generation to another by means of education. They see the school as aiming essentially at the transmission of the basic values derived by comprehensive philosophic study and hence they see in educational philosophy the basic source from which objectives can be derived.

The point of view recommended is that no single source of information is adequate to provide a basis for wise and comprehensive decisions about the objectives of the school. Each of these sources has certain values to commend it. Each source should be given consideration in planning. In this way educational programs may be developed that are flexible and suitable for any specific public school situations irrespective of whether the situation is influenced primarily by only one or any combination of these varying points of view concerning educational objectives.

While the objectives in these volumes contribute to solving the difficult problem of delineating a curriculum, they should not be considered as a final and perfect product. Any set of objectives may in fact be considered tentative, requiring continuous updating and reevaluation to the educational purposes and programs at hand. To have critical comments made about one's objectives should be taken as a compliment, since this can only happen when one has taken the trouble to think them out and write them down.

In spite of the great effort and man-hours that have gone into this task of compiling the objectives in these volumes, a number of the objectives listed cannot yet be considered to be "true" objectives (if by objectives we mean instructional outcomes described in performance terms). In fact, the editors wish to make the following critical comments as to some of the reasons why some of the objectives herein contained are open to multiple interpretation.

1. Some describe a classroom activity taking place during the process of learning, rather than the performance to be exhibited by the proficient student after learning.
2. Some lack a description, or even a suggestion of, the stimulus conditions under which a student is to perform. Conversely (and perversely), stimulus conditions are occasionally included when seemingly unimportant.
3. Some statements (I use that term rather than objectives) fail to suggest any sort of criteria. Though all objectives do not demand criteria, this lack, perhaps more than anything else, makes for vagueness.

The objectives in these volumes are the objectives for Project PLAN with slight editorial and organizational modifications. Project PLAN is a system of individualized education operative at grades one through twelve in the subject areas of language arts, mathematics, science, and social studies. Project PLAN was conceived by Dr. John C. Flanagan and to an extent evolved from the findings of Project TALENT, a large-scale, long-range project involving the collection of comprehensive information about education in the United States. Project TALENT involved the testing of a sample of 440,000 students in 1,353 secondary schools in all parts of the country in March, 1960, with subsequent follow-up studies. Through Dr. Flanagan's efforts, Project PLAN was brought into being in February, 1967, as a joint effort of the American Institutes for Research, Westinghouse Learning Corporation, and thirteen school districts.[1] Dr. Flanagan has continued to direct the developmental and research work on Project PLAN since that date and is an editor of these volumes. Assisting in the developmental work of Project PLAN has been Dr. Robert F. Mager. Dr. Mager is well known for his book, *Preparing Instructional Objectives,*[2] and his philosophy was followed in the development of the objectives in these volumes, of which he is an editor.

The cooperating school districts furnished classroom teachers each year from 1967 through June 1970 who developed the objectives and prepared the Teaching-Learning Units to accomplish the objectives under the supervision of American Institutes for Research and Westinghouse Learning Corporation professional personnel. The director of these activities was Dr. William M. Shanner, the third editor of these volumes. The teachers, at the end of each year, returned to their respective school districts to initiate the instructional programs organized from the objectives.

1. Archdiocese of San Francisco, Department of Education, San Francisco, California; Fremont Unified School District, Fremont, California; San Carlos Elementary School District, San Carlos, California; San Jose Unified School District, San Jose, California; Santa Clara Unified School District, Santa Clara, California; Sequoia Union High School District, Redwood City, California; Union Elementary School District, San Jose, California; Bethel Park School District, Bethel Park, Pennsylvania; Hicksville Public School District, Hicksville, New York; Penn-Trafford School District, Harrison City, Pennsylvania; Pittsburgh Public Schools, Pittsburgh, Pennsylvania; Quincy Public Schools, Quincy, Massachusetts; Wood County Schools, Parkersburg, West Virginia.

2. Mager, R.F. *Preparing Instructional Objectives.* Palo Alto: Fearon Publishers, 1962.

The objectives in these volumes, then, have originated from teachers and have been tried out in schools. I wish to acknowledge the efforts of those teachers who were assigned by their school districts to work a year at the American Institutes for Research in Palo Alto, without whose contributions the objectives in these volumes would not have been possible.

Archdiocese of San Francisco, Department of Education: Sister Maura Cole, Marian Bonnet, Janice Edminster, Sister Charlene Foster, Sister Bernice Heinz, Sister Patricia Hoffman, Sister Mary Vincent Gularte, Sister Anita Kelly, Sister Jeanne Marie Sosic

Bethel Park School District: Lora Moroni, Gordon Lepri, James Johnson, Judith Andrews, Flora Belle Faddis, David Loadman, Mary Lou Ertman, Roger Johnson, Robert N. Manson, Anna Marie Kerlin, Frances Chase, Robert M. Caldwell

Fremont Unified School District: Lyndall Sargent, Gail Pagan, Rex W. Estes, Caroline Breedlove, Monique Lowy, Charles Swanson, Eileen Trefz, Robert Fairlee, Beverly Ulbricht, Forrest W. Dobbs, Roy C. Fields, Bertram K. Robarts

Hicksville Public School District: Elayne Kabakoff, Richard C. Leuci, Terrence Boylan, Janet Findlay, Willard Prince, Edward Albert, Phyllis A. Kabakoff, Lawrence Dauch, Gerald Shanley, Marjorie Giannelli, Tom Bannan, Gerard F. Irwin

Hughson Union High School District: Warren Green

Penn-Trafford School District: Gary Fresch, Mary Ann Kovaly, Michael Demko, Jack Reilly, Victor Bohince, David Garvin, LaVelle Hershberg, R. Bruce Robinson

Pittsburgh Public Schools: Ann Mulroy, Jean Brooke, Kenneth Fraser, Shirley Fullerton, Ruth Aaron, Donald Coudriet, Cecilia Sukits, Carmen Violi, Samuel D. Martin, Paul J. Schafer, Mary South, Patricia Sellars

Quincy Public Schools: Jean Ann MacLean, Priscilla A. Dauphinee, Francis Keegan, Katherine Norris, Dennis Carini, Richard Russell, Stephen Fishman, Jack K. Merrill, Marcia A. Mitchell, Robert J. Mattsson, Margaret E. Flynn

San Carlos Elementary School District: Helen Dodds, Natalie Klock, Edith Bryant, Maxine Ross, Elizabeth Movinski, Martha A. Elmore, Charles B. Whitlock, Betty Lee, Lee Jensen

San Jose Unified School District: Allaire Bryant, Rise Berry, Hal Garrett, Kathy Roberts, William Harvel, Judy Opfer, Judi Wells, Don Crowell, Oran T. Adams, Marilyn D. Johnson, Alice S. Anderson, Sylvia Atallah

Santa Clara Unified School District: Nancy Wylde, Ruth Hessenflow, Arthur A. Hiatt, Herman Neufeld

Sequoia Union High School District: Gale Randall, Rex Fortune, Robert W. DuBois

Union School District: Jo Ann Risko, Peggy Schwartz, Rose Yamasaki, Glenn Moseley, Sue Coffin, Tod Hodgdon, Barbara S. Donley, Frank Kelly

Wood County Schools: Roberta Adkins, Mary Rector, Larry Myers, Virginia Haller, John Hoyes, Connie Chapman, Ada Ardelia Price, David V. Westfall, Nancy M. Rice, John W. Apgar

In addition, the contributions of the following persons should be acknowledged: Mary June Erickson, language arts; Josephine Matthews, Dr. Marie Goldstein, and Dr. Gordon McLeod, mathematics; Marvin Patterson, science; Dr. Vincent N. Campbell, social studies; Sarah Russell, primary; Katheryn Woodley, Dr. Mary Willis, Debbra Michaels, performance standards; and Dr. Helen Dell, editorial.

Final acknowledgment should go to those who use the objectives in these volumes. Objectives alone, an educational program, they do not make. They provide at best only a framework. The responsibility for the learning must rest on the student, guided by the teacher, and supervised by the school administration.

William M. Shanner

Palo Alto, California
December 15, 1970

INTRODUCTION

The PLAN Master List of Objectives has been prepared as a reference book for PLAN teachers and administrators. The book is divided into three sections. Objectives in the first section are for primary level concepts and skills, in the second section for intermediate level and in the third for secondary level. The objectives in each section of the book are organized into one of four subject areas, mathematics, science, social studies and language arts. Many of the objectives are found in more than one subject area, and these are indicated by an asterisk.

Each subject area has been subdivided into major concept sections. Within these sections are terminal and transitional objectives. Terminal objectives are defined as major growth points in the cognitive, skill and affective development of students. Educators can specify terminal objectives as the ones they wish their students to achieve at the end of a definite time block. Such a time block might be at the end of third grade, at the end of eighth grade and at the end of the high school experience.

Transitional objectives are listed under each terminal objective. They are defined as short term behavioral objectives, that is, concepts and skills to learn as prerequisites to the achievement of a terminal objective. A student may spend six months or several years achieving a series of transitional objectives before he is ready to challenge the terminal objective. Transitional objectives are organized sequentially under a terminal objective, if sequence is important, and are clustered when they have a common theme. It is not necessary to achieve every transitional objective before challenging the terminal objective. Since many transitional objectives are cross-referenced in several concept areas, students may achieve transitional objectives which will simultaneously support several terminal objectives.

Each objective has been written to prescribe one of six designated levels of performance The performance levels, based on Bloom's *Taxonomy of Educational Objectives*[1] are indicated by the roman numerals at the end of each objective, as in the following examples:

2372 After reading a fictional selection at the appropriate reading level, predict future consequences. (III)

5045 Analyze a given selection by inferring the author's intent and by drawing conclusions from the evidence presented. (IV)

The verbs used in the objectives are standardized to each performance level as defined in the glossary of this book. Some of the objectives are worded differently than they are in the present TLU's to correspond with the revised performance level and standardized verb list.

1. Benjamin S. Bloom, *Taxonomy of Educational Objectives*, David McKay Company, Inc., New York, 1956.

PERFORMANCE LEVELS

LEVEL I This level requires only memorization of factual information or major topic headings. Questions on this level would not require opinions or interpretation of facts. An example of level I would be:

0352 Identify the following properties of animals: how they eat, how they grow, how they change, how they move by themselves, and how they have babies. (I)

The following verbs are used in level I objectives:

Language Arts, Social Studies, Science

answer questions	locate
copy	list
define	match
finish	name
follow directions	pronounce
identify	reproduce
indicate	select
label	spell
	tell (retell)

Mathematics

copy	list
finish	match
identify	reproduce
	tell

LEVEL II Objectives designated as level II require the comprehension of information or the use of a skill in a different context from the original. Students may be required to classify, describe, or interpret information. If a student is required to recognize information, he knows that the suggested answers will not be taken verbatim from the original material but will be given in a different context. This is in contrast to the use of the verb "identify" in level I which indicates a memorization of facts or concepts. A student may also be asked to draw conclusions or summarize information in level II. Tasks in mathematics at this level include solving numerical problems and algebraic equations and making graphs. If a formula is suggested for solving a problem, it is considered to be level II, differentiating the task from one where a choice of formulas or operations must be made. Writing sentences with the appropriate sentence structure or grammar usage is also considered level II. An example of this level would be:

4940 Explain what is meant by adapting to environment (both biologically and culturally), and cite two examples of races adapting to their environments. (II)

The following verbs are used in level II objectives:

Language Arts, Social Studies, Science

classify	order
complete	read
construct	recognize
describe	suggest
draw conclusions	summarize
explain	use
express	relate
interpret	rewrite

Additional verbs used only in Mathematics and Science at this level:

add	integrate
calculate	measure
conclude	multiply
construct	organize
count	plot
define	put in order
derive	record
divide	represent
estimate	simplify
expand	solve problems, equations
factor	square
find	subtract
graph	test
illustrate	translate
infer	verify
	write numerals

Objectives designated in levels III to VI require a higher cognition level than the two previous levels. Intermediate and secondary terminal objectives are found principally in these higher categories. These levels may require a certain amount of memorizing or simple comprehension but the end result should be a more complex cognitive performance.

LEVEL III Objectives indicated as level III performance require the student to make an application of a principle, concept or skill. The student will be required to choose from several possible principles, formulas or concepts to demonstrate this performance level. The writing to be completed for objectives designated as level III does not require a creative effort. Writing is used instead to demonstrate an understanding of ideas, to support or refute a solution to a problem or to prepare an oral presentation. Objectives which require the ability to predict consequences and employ experimental procedures in finding solutions to problems designate level III performance level also. An example of this level would be:

8415 Present evidence from world history to support or refute this statement: *Too much involvement in foreign affairs over too long a time weakens a nation internally.* (III)

The following verbs are used in level III objectives:

Language Arts, Social Studies, Science

act out	participate
apply	predict
communicate	prepare and present
debate	present
demonstrate	pretend (role playing, perform)
discuss	support or refute a solution
find (information)	take notes
keep records	write
make, draw	

Additional verbs used only in Mathematics and Science at this level:

approximate	prove
determine	select
differentiate	solve word problems, problem situations
evaluate	tabulate
perform	write equations, problems, number sentences

LEVEL IV Objectives indicating level IV performance require an organization or analysis of ideas in a far more complex manner than the lower performance levels. Analyzing an author's writing involves making inferences about the author's interest or convictions, his freedom from bias or the validity of his arguments. Some objectives of this level require students to distinguish facts from hypothesis and factual statements from normative statements. The analysis of elements to show relationships and the forming of generalizations are cognitive skills which are also a part of this performance level. An example of level IV would be:

5256 Using the mass media as resources, analyze two or more viewpoints on a controversial issue. (IV)

The following verbs are used in level IV objectives:

Language Arts, Social Studies, Science, Mathematics

analyze	form generalizations
determine	infer
differentiate	organize

LEVEL V Students are asked to use their creative skills when achieving objectives designated as level V. They may combine and organize ideas in a unique way, design a plan for solving a problem, develop a new formula, or write an original composition. This requirement is distinguished from the writing in level III objectives by the creativity it involves in contrast to the reporting of facts and observations. Many of the secondary terminal objectives require this level of performance. An example of level V would be:

5269 Design, set up, and perform an experiment that will demonstrate that there is a 2:1 hydrogen to oxygen ratio in water. (V)

The following verbs are used in level V objectives:

Language Arts, Social Studies, Science, Mathematics

combine and organize	produce
design	write (original composition)
develop	

LEVEL VI Performance level VI requires the most complex skills and conceptualization of ideas in the evaluation of plans, procedures, techniques or solutions to problems. Evaluations at this level are made on the basis of specific criteria. They cannot be made without a thorough consideration of all the facts and of the effect that ideas may have on efficiency, economy, utility and human problems. Terminal objectives requiring this performance level are supported by transitional

objectives requiring the analysis of ideas, solutions to problems, and procedures. Although evaluation objectives are of the highest cognitive level, they may also be transitional objectives which enable the students to select the most appropriate ideas or techniques to produce a creative work in a terminal objective. An example of level VI would be:

2374 After reading a book at an appropriate reading level, evaluate the validity of the message in terms of personal experience. (VI)

The following verbs are used in level VI objectives:

Language Arts, Social Studies, Science, Mathematics

compare and contrast
evaluate
make judgments

Educators can use the PLAN Master List of Objectives as a reference book in the evaluation of goals for their students. The index provides a reference from module to concept organization. The modules are listed in the index sequentially by number. The objective numbers refer to page numbers where each can be found. The reader will then be able to find the terminal objectives which each transitional objective supports.

Helen D. Dell, Editor